Measuring Advice-Giving Skill: Evidence from a Large Online Experiment

[Click for the latest version]

Ashley Perry
New York University Abu Dhabi

Ben Weidmann
University College London

David Deming
Harvard University

Abstract

Open-ended advice is ubiquitous, yet we know little about who is good at giving it. We introduce a method for measuring the causal effect of individual advisors on the performance of the people they advise. In a preregistered online experiment, 157 advisors learned hidden information about an abstract task and provided audio advice to 64 advisees ($N = 10,048$). We find that a one SD increase in advisor quality raises advisee performance by 0.05 SD (p=0.015) and that advisor effects predict the performance of held-out advisees. This 'good advisor' effect size is twice as large as the difference in task performance between college-educated and other advisees. We find that being good at the task is unrelated to being good at advising on it. Moreover, advisees' impressions are informative: more effective advisors are rated as sounding more trustworthy. Methodologically, we find that LLM rubric scores of advice quality are positively correlated across two diverse advice-giving tasks, but we detect no relationship between rubric scores and advisors' causal impact. Our results highlight the value of outcome-based measures of advising skill. Substantively, we use our approach to explore gender heterogeneity and find evidence that the quality of advice matters more for mixed-gender pairs.

JEL Codes: D23, D83, D91

Keywords: advice, management, qualitative data, performance, gender

This version: October 2026

Acknowledgments. We thank Tommaso Batistoni, Ernesto Reuben, and CBID seminar participants for their helpful comments and suggestions. We also thank Muhammad Abdullah for excellent research assistance. The project received ethical approval from the NYUAD HRPP Office (HRPP-2026-55). A randomized controlled trials registry entry is available at https://www.socialscienceregistry.org/trials/18830.

Advice is a recommendation about what to do, feel, or think in response to a problem (MacGeorge & Van Swol, 2018a). The type of advice communicated can range from closed-ended recommendations (Sniezek & Buckley, 1995), to more open-ended language (Blunden et al., 2025). In general, advice is important as it facilitates knowledge transfer in a range of different contexts such as offices, schools, and financial markets (Brooks et al., 2018; Emanuel et al., 2026; Sandvik et al., 2020, 2026; Shen et al., 2025; Yin et al., 2025).

Clearly it is beneficial in these various settings to identify which individuals are good or not at communicating advice. However, open-ended advice presents a challenge in this regard because it is inherently unstructured, making it difficult to accurately quantify its impact at the individual level. Indeed, in many empirical studies with open-ended advice one can answer questions at the aggregate level, such as comparing advice to no-advice, the effects of different training programs on the provision of advice, or if the content of advice varies across groups (Ambuehl et al., 2025; Bucher-Koenen et al., 2025; Dimitriadis & Koning, 2022).

One potential approach to evaluate advice at the individual level are qualitative measures such as surveys (Feng & MacGeorge, 2010). For example, a meta-review of mentoring-assessment studies in medicine found that questionnaires and surveys were the dominant assessment method (Ng et al., 2020). Moreover, research on advice commonly measures recipients' perceptions of advice quality and intentions to follow it (Vu et al., 2025). Although these measures allow for individual assessment of advice they are often process-based and not clearly linked to actual implementation or downstream outcomes of advice. This makes it difficult to evaluate individual advice-giving effectively.

In this paper we introduce a novel measurement protocol designed to quantify the causal contribution of open-ended advice on advisee performance. We use the method of repeated random assignment (Weidmann & Deming, 2021). Our method uses a 'sequence task'. This task has two important features that allow for the measurement of advice-giving skill: advisee performance is objective and there is hidden, structural information that advisors can provide to aid advisee performance. Importantly, in our design we measure *how* advice is given and not other aspects of advice, such idea generation. This is because we disclose this hidden information to advisors, so that they are all equally informed of the same crucial information. Hence, the quality of advice is not a consequence of how familiar an advisor is with the particular task they do.

We implement and test our approach using a large, preregistered online experiment. We recruit two types of participants: advisors ($N = 157$) and advisees ($N = 10,048$). Both sets of participants initially complete the sequence task uninformed about the hidden information. This provides our baseline estimate of task skill. We then disclose the hidden information to the advisors but not the advisees. Advisors then make an audio recording of their advice on how to approach the sequence task. We incentivize advisors by making their bonus a function of their advisees' post-advice performance. We argue that communication of this hidden information from one person to the next reflects an important aspect of advice-giving: the transfer of relevant knowledge from an informed advisor to an uninformed advisee. After advisees listen to the advice recording of their advisor, they complete two further sequence tasks.

We find that after controlling for both advisor and advisee baseline performance, an advisor

whose estimated performance is one standard deviation above the average improves advisees' performance by around 0.05 standard deviations ($p_{RI} = 0.015$).[1] This effect is robust to a variety of controls. Although our estimated effect size is modest, it is just over twice as large as the effect on advisee baseline task performance from having at at least a college education (0.02 standard deviations, $p = 0.089$). Moreover, we believe it is a reasonable size given the focus on a single skill. It is between a fifth and two-fifths the size of other work using repeated random assignment to measure broader skills such as team players or management skills (Weidmann & Deming, 2021; Weidmann et al., 2026). Importantly, advisor-effect estimates predict held-out advisee performance ($p = 0.010$). We find that an advisor's estimated performance or advice-giving skill is not predicted by being good at the task: as an advisor's baseline performance in the sequence task does not predict their advisees' post-advice performance. Similarly, we find that advice-giving skill is not associated with measures of fluid intelligence (Raven & Raven, 2003) or social skill (Weidmann & Xu, 2025).

We demonstrate the value of our approach to quantifying the performance of open-ended advice. As part of our preregistered design we developed a second advice-giving task for advisors to complete. In contrast to the main task, we ask advisors to engage in a naturalistic advice-giving exercise, based on a workplace scenario. The advisor plays the role of a mentor providing guidance to a junior colleague who has recently received a negative performance review. This task requires advisors to engage in a dynamic back-and-forth conversation with a voice-based LLM. While this approach has several clear advantages, an obvious downside is the lack of an objectively measurable outcome to assess the causal impact of advice. Because of this, we use process-based rubric evaluations, reflecting the dominant approach in organizations and the academic literature (see for example, Feng & MacGeorge, 2010; Heslin et al., 2006; Vu et al., 2025).

To examine the difference between 'process-based' and 'objective' scoring, we generated common rubric scores for both the sequence task, and the naturalistic task (using multiple reasonable rubrics). A robust pattern emerges: the rubric-based scores positively correlate across the two divergent advice-giving tasks — but these rubric-based scores are *not* correlated with our objective measure of an advisor's causal contribution to advisee performance. This finding underscores the value of outcome-based measures of performance when considering how to evaluate advice.

Last, we use our main measurement approach to explore the role of gender in advice-giving. This is because men tend to hold more senior positions within organizations (Benson et al., 2026; Blau & Kahn, 2017; Cullen & Perez-Truglia, 2023; OECD, 2025), and consequently advice likely flows more often from men to women (O'Brien et al., 2010). For example, in a 2025 survey of 1000 US companies, 49% of entry level roles were held by women but at manager level, the next rung on the ladder, this dropped to 41%, and this pattern continues all the way up to C-suite where only 30% are held by women (McKinsey & Company & LeanIn.Org, 2025). Hence, understanding the causal impact of advice on different gender pairings is important. We find that advice quality is more consequential for performance – both good and bad – in gender-mixed pairs of advisors and advisees. In particular that advice quality of male advisors is more

[1] This p-value is based on our preregistered randomization inference approach.

consequential for female advisees as compared to female advisors, 0.086 standard deviations ($p_{RI} = 0.021$). As well as the advisor gender being perceived through their voice, it is likely that other attributes are also instantaneously inferred from their voice (McAleer et al., 2014). Hence, we explore how advisees' perceptions of the advisor predict our objective advice performance measure. We find positive associations between an advisor's performance and the advisor's perceived competence, likeability, and trustworthiness.

Our paper makes three contributions to the literature. First, we build on literature that identifies – and quantifies the impact of – managerial skills (Hoffman & Tadelis, 2021; Minni, 2026; Weidmann et al., 2026). Our contribution is to develop a new tool focused on advice-giving skill. Our approach uses the method of repeated random assignment (Weidmann & Deming, 2021). This method often requires contemporaneous human-to-human interaction, which can constrain the number of assignments. We show that this approach can be used in asynchronous human-to-human interactions, which means that it has the potential to be scaled to yield more precise individual-level estimates.[2]

Second, we contribute to the vast literature on advice (see, for reviews, MacGeorge & Van Swol, 2018b; Schotter, 2023). A large body of work in psychology has studied advice through the judge-advisor system (JAS) where the judge is the recipient who decides on an action based on the recommendation of the advisor (Bonaccio & Dalal, 2006; Sniezek & Buckley, 1995). Often the focus is on the utilization of the advice by the judge, the extent to which they follow the advice they receive (Van Swol et al., 2018). A common finding is that judges tend to discount advice they receive (Yaniv & Choshen-Hillel, 2012; Yaniv & Kleinberger, 2000). A recent meta-review suggests that the perceived advice quality of the advisor is the biggest factor determining this discounting (Bailey et al., 2023). Advice quality in this context refers to the expertise of the advisor and not how effectively they communicate information. In our setting we inform recipients that their advisor has privileged information about our task, to attenuate this expertise channel, which allows us to study the quality of the communication itself. Other work studies factors that determine the extent to which advice is shared, such as the motivational effects of giving advice (Eskreis-Winkler et al., 2018), whether it conveys positive or negative information (Eskreis-Winkler & Fishbach, 2020), or the reputation concerns that seeking advice might signal at work (Heursen et al., 2026). We complement this literature by focusing on the quality of the advice given. Our workplace advice task has a similar methodology to situational judgment tests, where individuals are given a social scenario and must make decisions about what they would do in that particular situation (Christian et al., 2010; Lievens & Sackett, 2012). These tests typically use closed-response questions to determine behavior, we provide a useful extension of this approach through our open-ended interactive format. Finally, scholars have developed methods to evaluate advice based on its emotional support (Burleson, 1982; High & Dillard, 2012) or perceptions of message features (Feng & MacGeorge, 2010; MacGeorge et al., 2004, 2016). Our diverse advice tasks, one with ground truth and the other without, allows us to evaluate the value of these types of metrics, especially when knowledge transfer is an important goal of advice.

[2]This method has also been used with human-to-AI interactions which also relaxes the constraints (Weidmann et al., 2025).

Third, we contribute to experimental research on the transmission of qualitative information. Recent work examines how senders' incentives affect what they communicate (Casaburi & Heursen, 2026; Hüning et al., 2022; Serra-Garcia, 2026; Thaler et al., 2025). We do not exploit variation in incentives, but instead ask whether some people communicate complex information more effectively than others. The role of voice in the transmission of information has been studied before, but in the context of economic indicators and belief-updating rather than advice-giving (Graeber et al., 2026). Finally, the communication of open-ended unstructured information, whether advice or feedback, has been experimentally studied in a range of settings, from start-ups to financial peer-advice (Ambuehl et al., 2025; Benoît et al., 2026; Blunden et al., 2025; Chatterji et al., 2019; Van Swol, 2011). These studies often utilize one-off interactions for communicating information. Whereas we use repeated random assignment of open-ended unstructured advice to identify the causal impact of the individual giving advice.

The rest of the paper is structured as follows: Section 1 contains the details of the experiment, Section 2 contains our analysis to identify the causal contribution of an advisor, understand characteristics of good advisors, and study our diverse advice-giving tasks. Section 3 contains our exploratory results of our novel task examining gender and perceptions of advisors. Finally, Section 4 contains our concluding remarks.

1 Experimental Design

1.1 Overview of Experiment

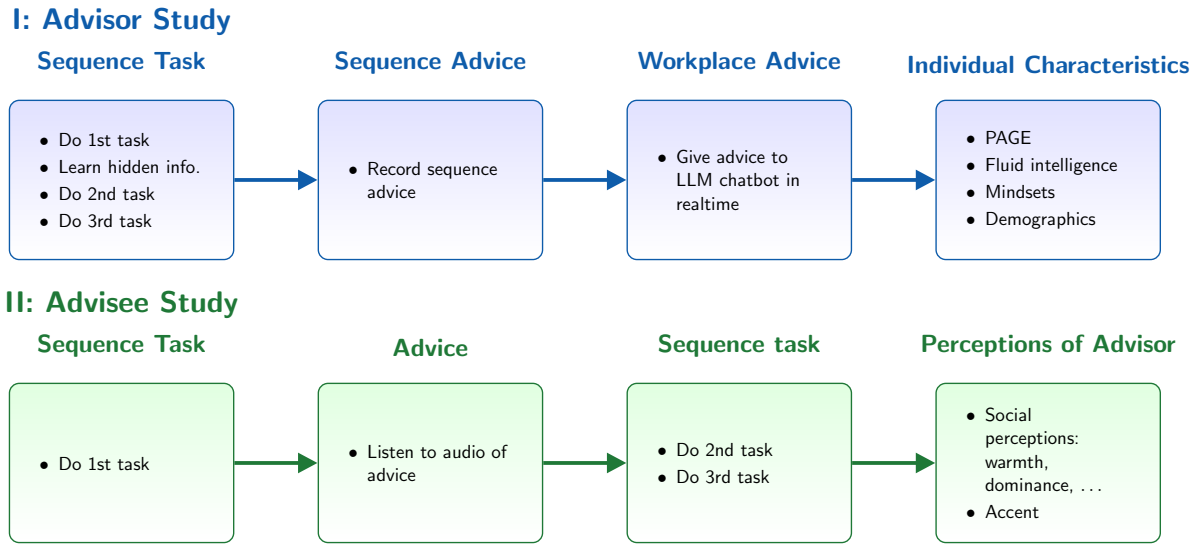
The experiment was preregistered at AEA RCT registry.[3] Participants in the experiment were assigned to one of two roles, an advisor or an advisee. The experiment consists of two studies, the advisor study and the advisee study. Figure 1 gives an overview of the two studies in the experiment.

The design utilizes a novel task called the sequence task. The sequence task has two features which allow us to measure the performance of open-ended advice. First, there is an objective performance measure. Second, there is specific hidden information about how the sequence task is constructed, knowledge of which can improve performance at the task. We leverage these features to mimic an important aspect of advice-giving, namely communication of relevant knowledge about a task or scenario from an informed advisor to a potentially uninformed advisee.

We recruited participants from the online platform Prolific and programmed the study in oTree (Chen et al., 2016) and the Harvard Skills Lab. We conducted the studies sequentially one after the other. First we collected data for the advisor study and after our preregistered data criteria checks were completed we then began data collection for the advisee study (see data criteria in Section B.2).

[3] AEARCTR-0018830, any deviations from the preregistration are mentioned in footnotes.

Figure 1: Experimental design flow chart



Note: The figure describes the flow of the experiment from the perspective of the participants. The experiment consists of two studies conducted sequentially. Participants were assigned to one of the two studies. The sequence task is described in more detail in Section 1.2. PAGE is an emotion perception task (Weidmann & Xu, 2025). Fluid intelligence is based on a Ravens matrix (Raven & Raven, 2003). Mindsets are beliefs about the malleability of people's intelligence and personality (Dweck & Leggett, 1988). Social perception are measures of personality attributes such as warmth and competence (Fiske et al., 2007; McAleer et al., 2014). For more information on these measures see Sections 1.4.1 and 1.5.1.

1.1.1 The Advisor Study

There are four sections in this part of the experiment. In the first section, participants, known as advisors, completed three sequence tasks. In the second section the advisors made an audio recording of their advice on how to do the sequence task.[4] In the third section advisors gave advice on a workplace vignette. This consisted of a real time back and forth audio conversation with an LLM based chatbot.[5] In the fourth section we measured individual advisor characteristics, including fluid intelligence (Raven & Raven, 2003), emotional perceptiveness (Weidmann & Xu, 2025), beliefs about mindsets (Dweck & Leggett, 1988), as well as demographics and other study related questions.

1.1.2 The Advisee Study

This part of the experiment took place after all data had been collected in the advisor study. We recruited a new set of participants, known as advisees. We randomized the allocation of advisor-advisee pairs, each advisor was allocated 64 advisees but each advisee was only allocated a single advisor. There are four sections in this part of the experiment. In the first section advisees completed the first sequence task. In the second section they listen to the sequence task advice of their assigned advisor. In the third section they completed the second and third sequence task. As each advisor was assigned a large number of advisees we randomly sampled advisees to undertake the fourth section, where they answered a series of questions about their perceptions

[4] We recorded this audio using software from https://voiceform.com.
[5] We conducted this audio interaction with software tools from https://www.skillslab.dev.

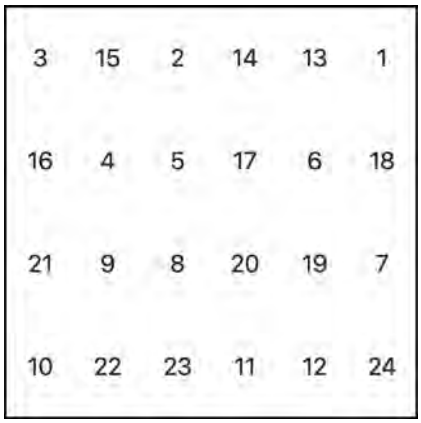
of their advisor.

1.2 Sequence Task Advice

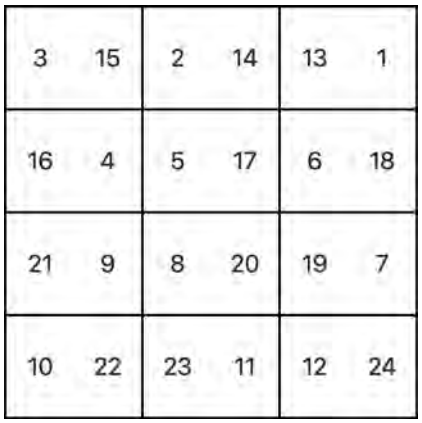
We developed the sequence task in order to leverage the method of repeated random assignment to capture the causal contribution of open-ended advice (Weidmann & Deming, 2021). As mentioned previously, to apply this method the sequence task has two important features. First, there is information that can be communicated about what to do that is richer than closed-ended advice e.g. "Choose option X" (Bonaccio & Dalal, 2006). Second, there is an objective metric to evaluate performance.

The sequence task is adapted from Der Zahlen-Verbindungs-Test (ZVT) or the number connection task (Oswald & Roth, 1978). In a typical ZVT participants are shown an image containing a sequence of randomly placed numbers from 1 to N. They click on the numbers in increasing order and their performance is measured by the time it takes to complete the sequence. We changed two features of the task to allow us to experimentally study open-ended advice. First, we change the performance metric to be the highest number a participant clicks to in a fixed amount of time. Second, we use a grid structure, hidden from the perspective of the participant, to allocate the numbers. Figure 2a shows an example of a sequence seen by a participant. Figure 2b shows the same sequence task with the hidden grid structure revealed. Within a cell of the grid the placement of numbers is random, but across cells there exists a placement rule which makes it easier to predict where the next number in the sequence is. We use the same three sequence tasks in both our advisor and advisee studies, two have a sequence length of 36 and one of 24. We also fix the time for each task to be 20 seconds. This paucity of time means that knowledge of the hidden grid structure is valuable as it can help a participant score highly (for more details see Section A.1 of the Appendix).

Figure 2: An example sequence task



(a) Grid structure hidden



(b) Grid structure revealed

Notes: Panel (a) shows a sequence task of length 24. Panel (b) shows the same sequence task as in panel (a) but with the grid structure revealed and consequently the placement pattern that the numbers follow. In this example, starting with 1 in the top right cell and snaking downwards and side to side until 12 in the bottom right cell. This pattern is repeated from 13 until 24.

1.3 Workplace Advice

Given the abstract setting of our sequence task advice we developed a second more natural advice-giving setting. We chose a workplace scenario as this is a setting in which people's advice is consequential to performance. All participants read the same vignette about a colleague called Casey, who has received their performance review which contained unexpected negative feedback. The colleague is seeking advice about what they should do before a follow up review. The participants also read a common set of expert recommendations before they engage in the voice conversation with Casey. Providing the same expert recommendations to all advisors helps to isolate communication skill rather than other skills such as idea generation. We used a male and female voice for Casey. This was assigned adaptively so that gender assignment would be balanced.

The stimuli for the vignette is drawn from real letters published on 'Ask a Manager' (Green, 2021). The expert recommendations are drawn on academic literature on workplace feedback (Gnepp et al., 2020). The prompt for the LLM contained these background details as well as general guidelines on how to respond, otherwise there were no restrictions. The latency of the responses of the LLM were minimal as we used the Realtime API of OpenAI which works directly with audio rather than converting audio to text or vice-versa (for more details see Section A.2 of the Appendix).

1.4 Advisors

Advisors completed three sequence tasks. They do the first one without any information about the hidden structure (see Figure 2a), this gives a baseline measure of their task skill. Then they were provided with detailed information along with annotated images about the hidden structure (see Section A.1 of the Appendix). Providing all the advisors with the same information about the hidden structure helps separate out communication skill from other skills such as problem solving.[6] Following the information disclosure of the hidden structure the advisors completed the second and third sequence task. After all the sequence tasks are completed they make a voice recording with their advice on how to approach the sequence task. The only restrictions they had were that the recording length could be up to two minutes and that they would be disqualified from the study if they revealed personally identifying details or offensive language. Advisors were informed that their advice would be shared with a random group of other Prolific users, who had not been informed about the hidden structure. Next advisors moved on to the workplace advice section where they engaged in an audio conversation with Casey, the LLM chatbot. This section was delivered through the Harvard Skills Lab in order to utilize the frontier LLM model for voice to voice interactions. Finally, after the conversation advisors then were re-directed back to the original oTree study website and then completed the final section of the study.

[6] As mentioned in Section 1.3, we also did this in the workplace advice setting to try to isolate the communication aspect of skill.

1.4.1 Individual Measures of Advisors

In the final section of the advisor study we measured a number of individual characteristics of participants that might be associated with advice-giving skills, such as emotion perceptiveness, fluid intelligence, and mindsets. We used PAGE (Perceiving AI Generated Emotions) to measure emotion perceptiveness (Weidmann & Xu, 2025), as it has been shown to be predictive of performance in environments where communication and motivation is important (Weidmann & Xu, 2025). PAGE captures emotional perceptiveness by presenting participants with a photo of a face expressing a particular emotion. For each face participants are asked to choose which word out of six best describes the expressed emotion. The canonical version of the test is called RMET (Reading the Mind in the Eyes test) Baron-Cohen et al. (2001). We used PAGE as it has a number of advantages over RMET: it has a more ethnically diverse set of stimuli and the stimuli have a larger emotional range. Fluid intelligence was measured with a set of Ravens Advanced Progressive Matrices (Raven & Raven, 2003), a commonly used assessment of the ability of solve novel problems. Implicit theories, also known as mindsets, are beliefs about the malleability of people's attributes, such as intelligence or personality (Dweck & Leggett, 1988). Advice has been shown to be more effective when the giver has a growth mindset (VandeWalle et al., 2001). We measured mindsets about others intelligence and personality using a Likert response scale. See Section A.3 of the Appendix for further details on all the individual measures.

1.5 Advisees

Advisees completed the same three sequence tasks in the same order as the advisors. Similarly to the advisors they completed the first sequence task without any information on the hidden structure, this gives a baseline measure of their task skill. Then each advisee was randomly allocated audio advice from an advisor, they heard the audio recording only once. We informed the advisees that their advisor had been given access to special information about the sequence task and that their advisor's bonus could depend on their performance. Subsequent to the advice the advisees completed the second and third sequence task. Next, a random sub-sample were asked a number of questions about their advisor. All advisees at the end answered a number of exit questions.

1.5.1 Advisee Perceptions of Advisors

As people readily form personality impressions during interactions, we measured advisees' perceptions of their advisor's personality attributes across a number of dimensions (Fiske et al., 2007; McAleer et al., 2014). Perceptions of advisors have been shown to influence expected advice quality (Pulford et al., 2018) and recipients' decision-making (Meshi et al., 2012). We measured six attributes such as warmth and dominance using a Likert scale. As we had a large number of advisees per advisor (64) we randomly sampled advisees to answer these questions. The median number of advisees to answer per advisor was 12. Furthermore, advisees were asked their perception of the advisor's accent. See Section A.3 of the Appendix for further details on these measures.

1.6 Identification Strategy

1.6.1 Notation and setup

The experiment contains two distinct and mutually exclusive roles: advisors and advisees. Individuals assigned to the advisor role are indexed by $i = 1, \ldots, n_A$ and those assigned to the advisee role are indexed by $j = 1, \ldots, n_E$. Let S_{ij} denote advisee j's mean score across the second and third sequence tasks after hearing the advice of advisor i. Let X_i^A and X_j^E denote advisor (pre information disclosure) and advisee (pre-advice) baseline (first) sequence task scores.

Each advisor i was randomly assigned $n_{Ei} = 64$ advisees and each advisee j only had one advisor. Assignment was stratified by gender (Female, Male, Other) and baseline (pre-advice) sequence-task performance (low if baseline score is < 11 and high if it is ≥ 11). To ensure balance within each block, the advisee was randomly assigned among the advisors with the fewest previous assignments in that block, subject to an overall maximum of 64 advisees per advisor. Assignment continued until each advisor had 64 advisees. We stratified the assignment of advisees to ensure that post-advice performance was not primarily driven by the advisees' own baseline attributes. Nonetheless, even within blocks there is variation in baseline scores, to address this we estimate the following model

$$S_{ij} = \gamma_A X_i^A + \gamma_E X_j^E + \delta^\top \mathbf{C}_j + \epsilon_{ij}, \quad \text{with} \quad \epsilon_{ij} \sim N(0, \sigma^2), \tag{1}$$

where $\mathbf{C}_j$ is the vector of indicators for the stratification blocks.

1.6.2 Estimating Advisor Performance

The residual ϵ_{ij} in equation (1) can be seen as a measure of the advisee post-advice performance that adjusts for the baseline sequence task skill of the advisor-advisee pair. If advisors were only assigned a single advisee it would be hard to determine whether variation in ϵ_{ij} is due to unmeasured individual attributes such as advice-giving skill or from idiosyncratic advisor-advisee match effects. However, by randomly assigning advisor-advisee pairs, such that advisors have many advisees but advisees only have one advisor, we can estimate μ_i – defined as the average causal impact advisor i has on advisee performance after controlling for individual differences in baseline sequence task skill – by averaging the residuals over the advisor i's n_{Ei} advisees

$$\hat{\mu}_i = \frac{1}{n_{Ei}} \sum_{j=1}^{n_{Ei}} \hat{\epsilon}_{ij}. \tag{2}$$

In our framework $\hat{\mu}_i$ is an estimate of the causal contribution of advisor i to advisee performance, conditional on each advisor-advisee pair's baseline skill. The precision of this estimate increases in the number of randomly assigned advisees. We have 64 advisees per advisor, which is a much larger number than previous studies utilizing this method. However, $\hat{\mu}_i$ is still noisy at the individual level. Thus we focus on whether ϵ_{ij} are correlated within advisors, in particular do advisors have a consistent effect on their advisees, after controlling for the advisor-advisee baseline sequence task skill. To do this we fit a multilevel model

$$\begin{aligned} \hat{\epsilon}_{ij} &= \mu_i + \tau_{ij} \\ \mu_i &\sim N(0, \sigma^2_\mu) \\ \tau_{ij} &\sim N(0, \sigma^2_\tau) \end{aligned} \tag{3}$$

In model (3), $\hat{\epsilon}_{ij}$ is a vector of baseline-adjusted advisor-advisee pair ij performance, μ_i is a random advisor effect for advisor i, and τ_{ij} is residual error. The subscript i is included to indicate that this analysis examines variation at the level of individual advisors. $\hat{\sigma}_\mu$ is our estimate of the typical advisor effect, in other words, the impact on advisee performance of having an advisor who is one standard deviation above the average advisor with respect to advice-giving skill.

For inference, we test the null hypothesis that advisors have no effect on their advisees after controlling for each advisor-advisee pair baseline sequence task skill (i.e. $\mathbf{H}_0 : \sigma_\mu = 0$). Our preregistered inferential approach is to calculate p-values using randomization inference (for details see Section B.1 of the Appendix). As robustness checks, we also report other estimates of uncertainty using a Wald estimator and profile likelihood estimates.

1.7 Recruitment, Sample, and Incentives

We recruited participants from the online worker platform Prolific. Table 1 contains descriptive statistics for the two types of participant roles, advisor and advisee. Column (1) corresponds to the 157 advisors who satisfied our preregistered criteria for valid audio advice (see Section B.2 of the Appendix for detailed criteria). As per our preregistration we recruited a gender balanced sample of advisors. Advisors have a broad range of ages with 28.7% older than 45 years of age. The majority, 61.8%, have at least a college degree. The ethnic make-up of the sample is broadly representative of the US population, for example 59.2% are only white which is similar to the 2020 census figure of 61.6% (Jones et al., 2021). Although not all categories match the census, our sample has 5.1% identifying as only Hispanic or Latino which is much smaller than the 18.7% 2020 census figure. We required advisors to be native English speakers and 94.3% were born in the US. We asked a random sub-sample of advisees the best guess of an advisor's accent, assigning the modal answer to an advisor, 12.1% of advisors are perceived to not have an American accent.

Column (2) corresponds to the 10,048 advisees who satisfied our preregistered completion criteria (see Section B.2 of the Appendix for detailed criteria). As per our preregistration we recruited a gender balanced sample of advisees. Similar to advisors, the age range of advisees is broad with 32.3% older than 45 years of age and 65.5% having at least a college education. Although the country of residence of advisees is not restricted to only the USA, the distribution of ethnic groups is quite similar.[7]

The sequence task score corresponds to the highest number a participant reached in a sequence within the 20 second time limit. The performance of the sequence tasks were incentivized

[7]In our initial preregistration we planned to only recruit English speaking advisees residing in the USA, but after the initial wave of data collection it became clear that we would struggle to recruit our preregistered sample of 10,048. Hence we opened up the country of residence to the following other predominately English speaking countries: United Kingdom, South Africa, Canada, Australia, Ireland, and New Zealand. This change was documented in the AEA RCT registry.

for both advisors and advisees using a piece-rate structure, a fixed amount multiplied by their score. Advisors were informed that there was an equal chance that their bonus would be determined by one of their own or one of their advisees' sequence task scores. At the end of the study they were told the outcome of this 50:50 lottery. If their performance was selected, then from their three scores one was chosen uniformly at random to be their bonus. If an advisee performances was selected, then from their group of advisees one of the scores from a second or third sequence task was chosen uniformly at random. Advisors were aware their bonus could depend on their advisees performance before they gave their advice. Also, they knew that advisees would be informed this was the case and that an advisees' bonus was only determined by that advisee. An advisees bonus was a random choice between their three sequence task scores.

Participants who completed the advisor study were paid a £6 participation fee and their piece-rate was £0.05 times their selected bonus score, for a maximum bonus of £1.8. The small size of the bonus and that there is a 50% chance an advisor's payment depends on their advice, does raise the possibility that advisors didn't put effort into the task. This would compress the variance in advisor effort which would mechanically shrink the advisor effect. However, this would imply that any effect we detect is a lower bound. Participants who completed the advisee study were paid a $1 participation fee and their piece-rate was $0.05 times their selected bonus score, for a maximum bonus of $1.8.[8] Advisor data was collected from August 12$^{\text{th}}$ until the 14$^{\text{th}}$ 2026 as each submission needed time to be verified (see Section B.2 of the Appendix for detailed criteria). Advisee data was collected from August 15$^{\text{th}}$ until the 26$^{\text{th}}$ 2026 in ten sessions, to allow time for the large number of observations to be collected and verified (see Section B.2 of the Appendix for detailed criteria).

2 Results

This section presents our main preregistered results (deviations are indicated in footnotes).

2.1 Advisor Effect

We identify a small but robust advisor effect using repeated random assignment of advice. Table 2 contains the estimated advisor effects ($\hat{\sigma}_\mu$) along with the p-values from our preregistered randomization inference procedure, with Wald and Profile likelihood for robustness. We also present coefficients on the standardized baseline scores, which are used to condition advisee post-advice mean standardized scores in model (1). We can see that an advisor's baseline score predicts little of advisees' post-advice mean scores, whereas the baseline performance of an advisee is highly predictive. For example in column (1) the advisor coefficient is close to zero (-0.004, $p = 0.68$) whereas the advisee coefficient is large (0.694, $p < 0.01$). The latter implies that an advisee whose baseline performance was one standard deviation above the mean has a post-advice performance 0.694 standard deviations higher than the mean. This indicates that even though advice does matter in this context, initial task skill of the advisee – good or bad – is important. Standard errors for these coefficients are in parentheses.

[8] The use of different currencies was a result of different funding sources.

Table 1: Descriptive Statistics

	Advisors (1)	Advisees (2)
Age		
18–24 years old (%)	8.9	11.0
25–34 years old (%)	36.9	31.4
35–44 years old (%)	25.5	25.2
45–54 years old (%)	17.8	16.4
55–64 years old (%)	8.9	10.9
65–74 years old (%)	1.9	4.2
75 years or older (%)	0.0	0.8
Gender		
Male (%)	50.3	49.3
Female (%)	48.4	49.8
Other (%)	1.3	0.9
Education		
Less than high school (%)	0.0	1.1
High school degree/GED (%)	9.6	14.4
Some college (%)	28.7	19.1
College degree or higher (%)	61.8	65.5
Employment status		
Part-time (%)	16.6	15.7
Full-time (%)	53.5	52.3
Self-employed or small business owner (%)	17.2	12.6
Not employed (%)	12.7	19.4
Born in the USA (%)	94.3	37.1
Accent		
Standard USA (%)	80.9	–
Regional USA (%)	7.0	–
Foreign non USA (%)	12.1	–
Country of residence		
Australia (%)	–	2.6
Canada (%)	–	10.2
Ireland (%)	–	1.6
New Zealand (%)	–	0.5
South Africa (%)	–	4.3
United Kingdom (%)	–	40.0
United States (%)	100.0	40.7
Race and ethnicity		
American Indian or Alaska Native (%)	0.0	0.2
Asian (%)	8.9	11.8
Black or African American (%)	16.6	16.3
Hispanic or Latino (%)	5.1	2.5
Middle Eastern or North African (%)	1.3	0.9
Mixed or multiple races (%)	8.3	4.5
Native Hawaiian or Pacific Islander (%)	0.6	0.1
Other ethnic identity (%)	0.0	2.0
White (%)	59.2	61.6
Sequence task 1 score	10.8 (3.2)	9.5 (4.1)
Sequence task 2 score	13.5 (4.3)	12.4 (4.7)
Sequence task 3 score	13.2 (5.0)	11.2 (3.9)
Audio length (seconds)	40.5 (22.6)	–
Fluid intelligence (Raven)	5.4 (2.8)	–
Emotion perception (PAGE)	11.2 (2.4)	–
Personality mindset (Likert)	3.3 (1.1)	–
Intelligence mindset (Likert)	3.3 (1.1)	–
Number of Observations	157	10,048

Notes: Categorical variables are reported as percentages. Continuous variables are reported as means with standard deviations in parentheses. Likert indicates a numerical rating scale from 1 to 5. Race and ethnicity categories are mutually exclusive. A dash indicates that the variable was not collected or is not applicable to that sample.

Column (1) presents our preregistered results. The typical advisor effect, in other words the impact of advice from an advisor who is one standard deviation above average, is 0.048 standard deviations ($p_{RI} = 0.015$).[9] The estimated reliability of the advisor-effect measure is approximately 0.21, that is 21.1% of the variation in the estimated advisor scores represents between-advisor signal. Given that an advisor's baseline score predicts little of their mean advisee post-advice score but we find that advice causally contributes to performance, suggests that being good at the task itself is not associated with being good at explaining it. To contextualize this effect size, regressing the standardized baseline (pre-advice) task performance of advisees on a dummy for having at least a college education, yields a coefficient of 0.021 ($p = 0.089$). This suggests the impact of a typical advisor on performance is twice as much as impact of a college education on solving the novel sequence task. Moreover, previous work employing repeated random assignment procedures have found effect sizes of 0.129 standard deviations for team players (Weidmann & Deming, 2021) and 0.221 standard deviations for general management skill (Weidmann et al., 2026). In comparison our effect is between a fifth and two-fifths of the size. However, in both those studies participants engaged in a number of activities contemporaneously, allowing for a thick social interaction where participants could impact each other on numerous dimensions. Whereas, by design this study delivered audio advice asynchronously in order to focus on communicating expert-informed advice at the expense of the thickness of the social environment.

Columns (2)-(5) present our estimate with the addition of a number of controls for advice, advisee, and advisor characteristics.[10] Columns (2) and (3) show that our estimate drops only slightly to 0.046 ($p_{RI} = 0.023$) and 0.046 ($p_{RI} = 0.019$) with controls for advice features and advisee relevant controls respectively. In column (4) we see that a relatively larger portion of the advisor effect is absorbed by the addition of advisor controls, it reduces to 0.037 ($p_{RI} = 0.078$). Although this is to be expected given the relatively smaller number of advisors as compared to advisees, the addition of advisor-level dummy variables will mechanically absorb some of the advisor variation. Column (5) includes all the controls and the estimated advisor effect reduces by a third to 0.032 standard deviations and is no longer statistically significant at the 10% level for our randomized inference procedure ($p_{RI} = 0.132$).

As an additional robustness test, we assess whether estimated advisor effects predict per-

[9] The randomized inference p-value is our preferred measure of uncertainty, but we include other measures, Profile Likelihood and a Wald estimate, as robustness checks.

[10] Advice features[a] includes the following: a binary indicator if the advisor used a headset to record their advice, the length of their sequence task advice in seconds, the time in seconds advisors spent reading their information about the hidden structure, a binary indicator if the advisee had a technical issue listening to the advice e.g. their advice was paused or they changed their headset. Advisee information[b] includes the following: an indicator for ethnic group, a binary indicator for at least a college-level education attained, an indicator for country of residence, an indicator for employment status, an indicator for data collection session, an indicator for pointer type used e.g. mouse or touchscreen, an indicator for device type e.g. laptop or tablet, a binary indicator for if they failed the first round of understanding questions, and an indicator for if they opened a dropdown of the sequence task instructions before the first sequence task. Advisor information[c] includes the following: an indicator for ethnic group, a binary indicator for at least a college-level education attained, an indicator for which region they live in the USA, an indicator for employment status, a binary indicator for if they were born in the USA, a binary indicator for if they failed the first round of understanding questions, and an indicator for if they opened a dropdown of the sequence task instructions, their score on the emotion perception test PAGE, and their score on the fluid intelligence Ravens test.

Table 2: Estimates of the Advisor Effect

	Dependent variable: Advisee mean score (S)				
Variable	(1)	(2)	(3)	(4)	(5)
Advisor effect ($\hat{\sigma}_\mu$)	0.048	0.046	0.046	0.037	0.032
⟨Randomization inference⟩	⟨0.015⟩	⟨0.020⟩	⟨0.017⟩	⟨0.078⟩	⟨0.132⟩
[Profile likelihood]	[0.018]	[0.023]	[0.019]	[0.095]	[0.152]
{Wald}	{< 0.001}	{< 0.001}	{< 0.001}	{< 0.001}	{< 0.001}
Advisor baseline score	-0.004	-0.004	-0.002	-0.003	-0.002
	(0.009)	(0.009)	(0.009)	(0.009)	(0.010)
Advisee baseline score	0.694	0.694	0.643	0.694	0.643
	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)
Controls					
Advice features[a]	–	Yes	–	–	Yes
Advisee information[b]	–	–	Yes	–	Yes
Advisor information[c]	–	–	–	Yes	Yes
Observations					
Advisor	157	157	157	157	157
Advisee	10,048	10,048	10,048	10,048	10,048
adj-R^2	0.4545	0.4547	0.4760	0.4544	0.4762

Notes: The dependent variable is the mean advisees' second and third sequence task score, standardized to have mean zero and standard deviation one (first the scores were standardized individually and then the mean of the standardized scores was standardized). All models include the randomized stratification blocks as fixed effects. The standard errors are in parentheses and clustered at the advisor level. The Advisor effects are estimated using model (3). We report p-values using three different approaches to inference, the null hypothesis being tested in all cases that ($\hat{\sigma}_\mu = 0$). For details on our inference procedure see Section B.1 of the Appendix. Advisor and advisee baseline score are the sequence task 1 scores, standardized within role type to have mean zero and standard deviation one, from model (1). See footnote 10 for details on the following controls [a]Advice features, [b]Advisee information, and [c]Advisor information.

formance for advisees out-of-sample using four-fold cross-validation.[11] Within each advisor, the advisor–advisee observations are randomly partitioned into four folds. In iteration f, we estimate model (1) on the other three folds, calculate the training-sample advisor effect $\hat{\mu}_i^{(-f)}$ as in equation (2), and use the training-sample coefficients to predict the outcomes in the held-out fold. Let $\tilde{\epsilon}_{if}$ denote advisor i's mean residual in held-out fold f. After rotating through all four folds, so that every observation is held-out once, we estimate

$$\tilde{\epsilon}_{if} = \alpha_{\text{CV}} + \lambda \hat{\mu}_i^{(-f)} + u_{if}. \tag{4}$$

The coefficient λ measures whether an advisor effect estimated from 3-fold advisees predicts the advisor's performance among held-out fold advisees. We test $H_0 : \lambda = 0$ against $H_1 : \lambda > 0$. We present the results in Table 3. Columns (1) to (4) show the cross-validation analyses for the hold-out fold. These estimates are noisier than our main analysis as we are dropping a quarter of the advisees for each advisor and the estimates are derived from equation (2). Column 5 aggregates the four folds and shows that on average, advisor performance predict out-of-sample performance ($p = 0.010$). That is, the advisor effects we are estimating robustly predicts performance. We also find a small but positive mean pairwise Spearman correlation of the four held-out advisor rankings of 0.07, suggesting that the exact rankings of advisors does vary across the folds.

Table 3: Predicting advisor performance out-of-sample

	DV: mean residual performance in the held-out fold				
	Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Overall**
	(1)	(2)	(3)	(4)	(5)
Advisor performance CV ($\hat{\mu}_i^{-f}$)	0.089	0.316***	0.274**	0.012	0.193**
	(0.126)	(0.127)	(0.121)	(0.119)	(0.083)
Observations	157	157	157	157	628
adj-R^2	-0.003	0.027	0.019	-0.000	0.012

Notes: The dependent variable is the mean causal contribution of advisor i on advisee performance (see equation (2)) in the holdout fold of data. Advisor performance CV is defined using the remaining three folds of data. Standard errors are in parentheses and are clustered at the advisor level as the same advisor contributes observations to each fold. Significance levels are denoted by: *** $p < 0.01$, ** $p < 0.05$, and, * $p < 0.1$.

We have established that we can estimate a robust causal contribution of advice. Now we present some descriptive analysis of the text of the advice. We compare the frequency of words used by high and low performers using our estimated advisor effect from equation (2). We categorize high (low) performers as the top (bottom) 20% of the distribution, with rounding this corresponds to 32 advisors in each group. To classify characteristic words for these two groups we estimate the likelihood a word is used in our entire advice data set ($N = 157$) and then estimate the word-specific divergence in the two sub-groups; low and high (Monroe et al., 2008).[12] Then for each word word we take the difference of the high groups estimated likelihood

[11] This analysis uses our preregistered model (Column (1) of Table 2)

[12] We pre-process the text such as removing stop words and lemmatizing words (see details in Section B.3 of the Appendix).

minus the low groups, to give a score of the relative importance of that word for two groups. Where positive values indicate the word is more characteristic of the high group and negative values of the low group. Figure 3 contains the 20 most characteristic words of the two groups. The blue bars with positive scores correspond to the high group and the orange bars with negative values correspond to the low group. Firstly, each word is present in both groups of advice (for details see Section B.3 of the Appendix). This means that top and bottom advisors mention the hidden placement pattern and the time constraints. However, one pattern that seems to emerge is that high performing advisors are more likely to focus on the time constraint and the low performing advisors are more likely to focus on the grid structure. On inspecting the advice, it does appear that advisors in the high performing group give advice on how to use the hidden placement pattern e.g., 'scan ahead'. Whereas, low performing advisors only mention the presence of a pattern without much advice on how to use that knowledge. In should be noted that this analysis gives associations between words and advisor performance rather than causal effects of particular words, which would require a different experimental design.

Figure 3: Word frequency analysis

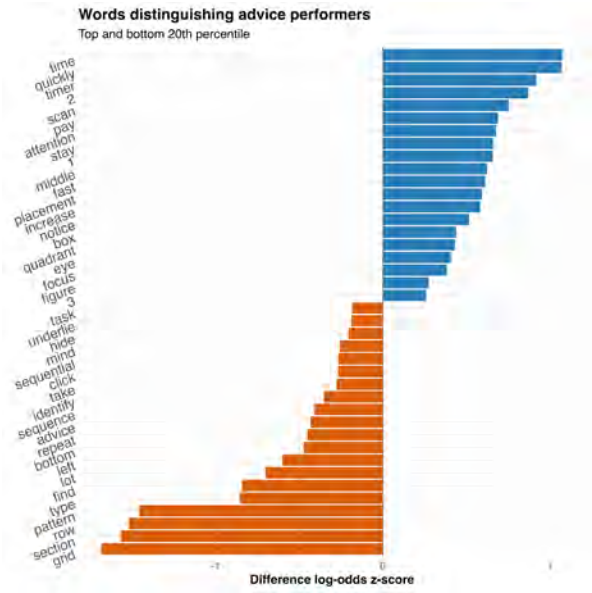


Figure 4

Notes: Plots the difference of the log-odds scores of lemmatized sequence advice text between the top and bottom 20% of the estimated performance ($\hat{\mu}_i$) distribution from equation (2). Positive values indicate words relatively more characteristic of top performers; negative values indicate words relatively more characteristic of bottom performers. Word frequencies were regularized using an informative prior constructed from all the sequence task advice using (Monroe et al., 2008). Words appearing in fewer than five top- or bottom-group documents combined were excluded. Bar length represents the magnitude of the standardized, prior-adjusted log-odds difference. The sample consists of all 157 advisors sequence task advice.

2.2 Characteristics of Good Advisors

We have demonstrated that we can causally identify advice-giving skill. We now explore what characteristics of advisors predict advice-giving skill ($\hat{\mu}_i$). Panel A and column (1) of Table

4 presents the correlates of advisor performance conditioning for the baseline advisor and advisee task skill against a number of advisor level attributes. Given the estimated reliability of our advisor estimate is 0.21, it is noteworthy that the maximum correlation with a perfectly measured variable cannot exceed $\sqrt{0.21} = 0.46$. We find no statistically significant predictor of advisor performance with other individual advisor characteristics. We do find small positive associations with advisor performance and advisor emotion perceptiveness ($r = 0.08$, $p = 0.34$) and fluid intelligence ($r = 0.07$, $p = 0.40$). For our mindset measures, where positive scores indicate a belief consistent with a more growth orientated mindset, we find small and negative associations with advisor performance for personality mindsets ($r = -0.04$, $p = 0.63$) and intelligence mindsets ($r = -0.12$, $p = 0.14$). This is in contrast to Heslin et al. (2006) who find a positive relationship with advice quality and the orientation of implicit theories i.e. a growth mindset. Although, in that study quality is not based on ground truth as it is in our context.

In Section 2.1 we showed that an advisor's baseline performance is not predictive of their advisees' post-advice performance. Together with the lack of correlations in this section it suggests that the advisor effect we identify may be a distinct skill.

Table 4: Correlates of advisor performance

	Advisor performance ($\hat{\mu}_i$) (1)	**Sequence advice rubric B** (2)
A		
Advisor attributes[a]		
Emotion perception (PAGE)	0.08	0.12
	(0.34)	(< 0.01)
Fluid intelligence (Ravens)	0.07	0.11
	(0.40)	(0.01)
Personality mindset	-0.04	-0.01
	(0.63)	(0.85)
Intelligence mindset	-0.12	-0.04
	(0.14)	(0.37)
B		
Advice evaluation rubrics[b]		
Workplace advice rubric A	0.04	–
	(0.60)	
Workplace advice rubric B	0.01	0.29
	(0.89)	(< 0.01)
Sequence advice rubric B	0.10	–
	(0.22)	

Notes: We report pairwise correlations of variables for all 157 advisors. All variables have been standardized to have mean zero and standard deviation one. In column (1) the dependent variable is the mean causal contribution of advisor i on advisee performance (see equation (2)) using our preregistered specification of model (1) i.e. advisor average causal contribution controlling for individual differences in baseline sequence task skill. In column (2) the dependent variable is the subjective evaluation of sequence advice based on common rubric B (for details on the rubrics see Section B.4 of the Appendix). [a]Advisor attributes see Section 1.4.1 for details. [b]Advice evaluation rubrics: See Section B.4 of the Appendix for details on rubrics. Workplace advice corresponds to the advice given in the workplace scenario (see Section 1.3) and Sequence advice corresponds to the sequence task advice (see Section 1.2). To express uncertainty, p-values are in parentheses. As stated in our preregistration our aim is to explore correlations and so we do not employ a multiple-hypothesis-test correction (Parker & Weir, 2022).

2.3 Advice-giving Skill Across Tasks

Next we examine the value of our approach to identifying the causal impact of advice by comparing it with no ground truth process-based measure. First we score an advisor's workplace advice using our preregistered evaluation rubric, which we denote rubric A (see Section B.4.1). This rubric has the following four dimensions: 'Acceptance', 'Future', 'Sustainability' and 'Responsiveness'.[13] Each dimension is scored on an integer scale. We use an LLM acting as a judge to score each advice text. The general approach is that we make two independent calls the API for the same advice text and prompt. If there is large enough disagreement on a particular dimension we make a third independent call to the API, blind to the previous two ratings and use the median rating as the score (for more details see Section B.4).

Panel B and column (1) of Table 4 contain the correlation of an advisor's workplace advice score based on rubric A with their advisor performance ($\hat{\mu}_i$). We find a small positive but not statistically significant correlation between these two different advisor level advice evaluations ($r = 0.04$, $p = 0.60$). This suggests that an advisor's performance in one advice setting does not translate to another.[14] However, our quantitative advisor performance ($\hat{\mu}_i$) is outcome-based, whereas rubric A measure is process-based and so it could be that they are measuring different aspects of advice.[15] We do not use rubric A to assess the sequence task advice as the prompt is specific to the workplace advice task. Instead we created a common rubric to both advice contexts, where we gave an LLM only the advice text from both tasks and asked it to develop a common rubric that could assess both types of advice, we denote this rubric B (see Section B.4.2). This rubric has the five dimensions 'Relevance', 'Actionability', 'Rationale', 'Feasibility', and 'Clarity'. Panel B and column (1) of Table 4 contain the correlations of an advisor's performance ($\hat{\mu}_i$) with their advice in both tasks evaluated using rubric B. We find small positive but not statistically significant association between advisor performance and the advice evaluated with rubric B for the workplace advice task ($r = 0.01$, $p = 0.89$) and the sequence advice task ($r = 0.10$, $p = 0.22$).[16] However, column (2) of panel B shows that the rubric based score positively correlates across tasks and is statistically significant ($r = 0.29$, $p < 0.01$). Figure 5 plots the individual scatter plots of these correlations.

Given the reliability of the advisor-performance estimate, it may be too noisy to provide adequate power. One method to address this is to utilize the advisee-level observations. We add the rubric B advice score into model (1) as a predictor variable and estimate how much of the advisor variance component is absorbed. In our preregistered model, the estimated variance component was 0.00233, corresponding to an advisor standard deviation of 0.048. Adding the rubric B reduced the estimated variance component to 0.00230, a proportional reduction of

[13] This is deviation from our preregistration, which has Warmth instead of Responsiveness, but given the back and forth nature of the interaction we feel this is a better performance dimension to measure as it captures the advisor's engagement with the LLM chatbot rather than just repeating stock phrases.

[14] Correcting for attenuation due to the estimated reliability of the advisor-effect measure ($\rho = .21$) gives an implied latent correlation of $r = 0.09$, with an approximate 95% confidence interval of $[-0.25, 0.43]$.

[15] We preregistered that we would explore the features of advice common to both settings. Although the specifics of this analysis was not detailed in our preregistration and so the analysis from here should be treated as exploratory.

[16] Correcting for attenuation for the workplace task gives an implied latent correlation of $r = 0.03$, with an approximate 95% confidence interval of $[-0.32, 0.37]$. The corresponding corrections for the sequence task are $r = 0.21$, with an approximate 95% confidence interval of $[-0.13, 0.55]$.

1.6%. The 95% confidence intervals for the proportional reduction are wide and included zero (−10.9% to 80.2%).[17] Thus, the point estimate suggests that rubric B does not explain much of the observed advisor heterogeneity, although the estimate is imprecise.

As a further robustness check we take the same prompt structure for rubric B but instead use evaluation criteria for advice derived from Advice Response Theory (Feng & MacGeorge, 2010), we denote this rubric C (see Section B.4.3). We find qualitatively the same pattern, there is no strong association between our outcome-based measure and the process-based rubric scores, but there is between the rubric evaluation of the two advice tasks (see Table B2 in Section B.4.3 of the Appendix). We also find similar results for the advisee-level exercise described above, a small estimated reduction in the estimated advisor variance component, 4.4% with wide 95% confidence intervals (−7.5% to 79.5%). These findings suggest the importance for ground truth outcome-based measures when assessing the quality of advice.

3 Exploratory Analyses

In this section we provide preregistered exploratory results with respect to our novel method for identifying the causal effect of advice. We focus in particular on gender and perceptions of advisor personality attributes.

3.1 Gender Heterogeneity

We know that men are more likely to be managers (Blau & Kahn, 2017; McKinsey & Company & LeanIn.Org, 2025), women are promoted less often (Benson et al., 2026; Cullen & Perez-Truglia, 2023), and in general there are large and persistent gender disparities in the workplace (Hospido et al., 2022; Hyland et al., 2020; Kleven & Landais, 2017; OECD, 2025). Given these well known gender differences in labor market outcomes, we stratified the advisee allocation by gender so that we could explore whether the causal impact of advice was moderated by gender (see details in Section 1.6.1). In particular, we are interested in the influence of advice in the four possible advisor-advisee gender groups: male advisor to female advisee, female advisor to female advisee, and so on.[18] These groups are mutually exclusive at the advisee level, but since advisors had both male and female advisees each advisor is in two of the four groups.

First, we partition our data into the four possible advisor-advisee gender groups: male advisor to female advisee, female advisor to female advisee, and so on. For each group we estimate the advisor effect using model (3), conditioning on advisor and advisee baseline performance.[19]

[17] Confidence intervals are calculated using 2,000 non-parametric bootstrap replications clustered at the advisor level. All bootstrap models converged. Between 2.8% and 3.9% of replications produced singular fits in which the advisor variance was estimated near zero. Consequently, intervals for the proportional reduction in advisor variance were strongly asymmetric because the baseline variance appears in the denominator.

[18] Participants could indicate a third option of "other" for gender. 1.3% of advisors and 0.9% of advisees chose this option. Due to the relatively small size of this group we drop these participants from the following analysis. Broadly speaking there are no large differences between genders within participant roles, although within advisees we find that females tend to perform slightly worse on the sequence task than their male counterparts (see Tables B3 and B4 of Section B.5 of the Appendix)

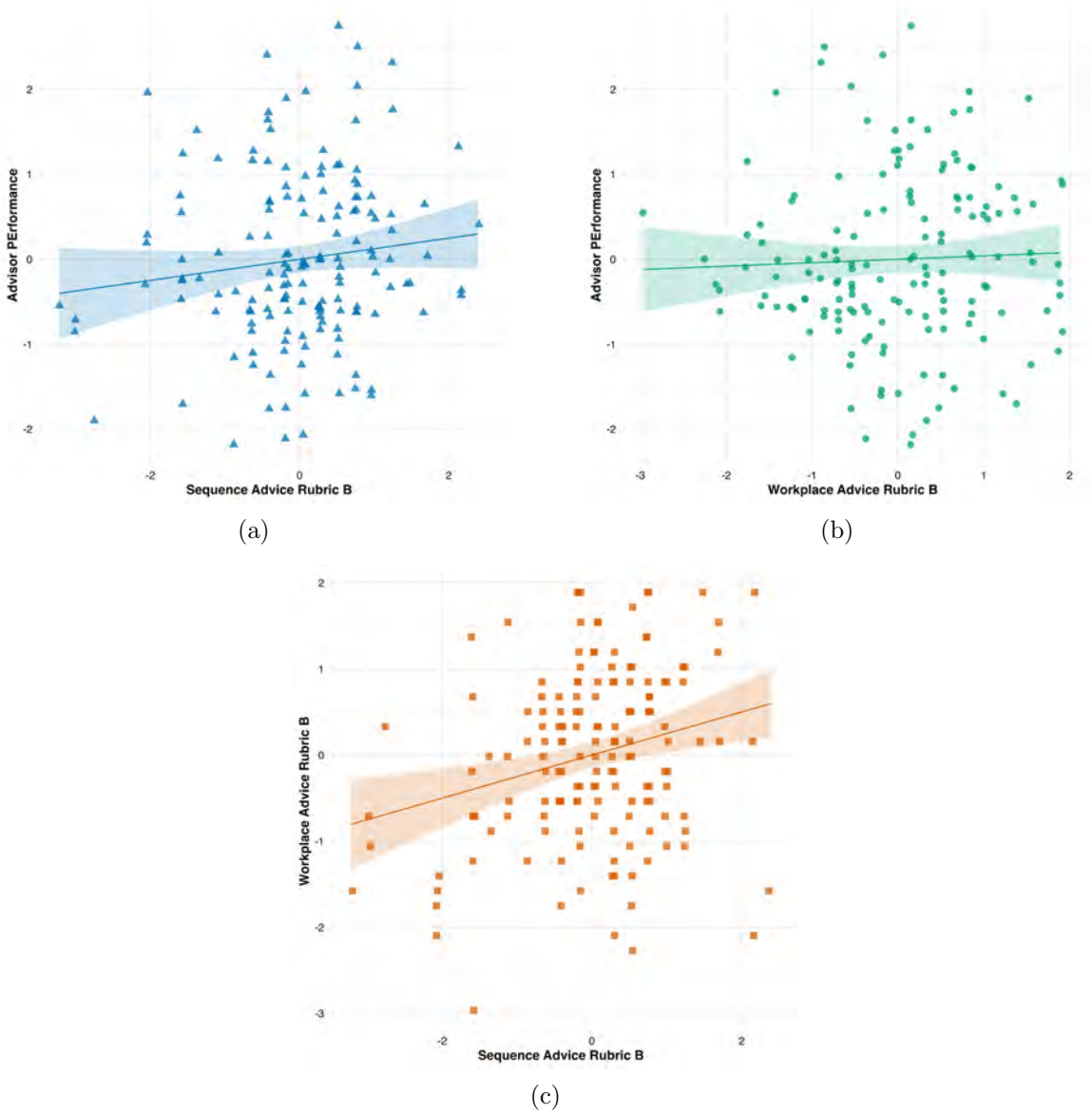
[19] We include only the performance randomized stratification blocks, low or high, since each sample only has one advisee gender.

Table 5: Gender Heterogeneity of Advisor Effect

	DV: Advisee mean score (S)			
	(1)	(2)	(3)	(4)
Advisor sample	Male	Female	Female	Male
Advisee sample	Female	Female	Male	Male
Advisor effect ($\hat{\sigma}_\mu$)	0.086	0	0.046	0
(Randomization inference)	(0.010)	–	(0.226)	–
Diff. advisor effects ($\hat{\sigma}^m_\mu - \hat{\sigma}^f_\mu$)	-0.086		0.046	
(Randomization inference)	(0.021)		(0.179)	
Observations				
Advisor	79	76	76	79
Advisee	2,514	2,428	2,397	2,494

Notes: The dependent variable (DV) is the mean advisees' second and third sequence task score, standardized to have mean zero and standard deviation one (first the scores were standardized individually and then the mean of the standardized scores was standardized). Advisors and advisees who indicated other as their gender have been dropped and only represent around 1% of either sample (see Table 1). All models include only the performance randomized stratification blocks as fixed effects. The columns represent different splits of the dataset. Column (1) are female advisees with male advisors, column (2) are female advisees with female advisors, column (3) are male advisees with female advisors, and column (4) are male advisees with male advisors. The Advisor effects are estimated using model (3). We report p-values using the random inference approach. For the advisor effect the null hypothesis being tested is ($\hat{\sigma}_\mu = 0$). For the difference of advisor effects the null being tested is ($\hat{\sigma}^m_\mu - \hat{\sigma}^f_\mu = 0$). For details on our inference procedures see Section B.1 of the Appendix.

Figure 5: Correlation of evaluation metrics



(a) (b)

(c)

Notes: Scatterplots of individual advisor performance between different performance metrics with OLS regression line and 95% confidence band. All variables are standardized to have mean zero and standard deviation one. Advisor Performance is the mean causal contribution of advisor i on advisee performance ($\hat{\mu}_i$) using our preregistered specification of model (1). Sequence Advice Rubric B and Workplace Advice Rubric B correspond to the subjective evaluation of an advisor's advice from that setting based on the common rubric B (for details on the rubrics see Section B.4 of the Appendix). Each panel contains the scatterplot of (a) Advisor Performance and Sequence Advice Rubric B, (b) Advisor Performance and Workplace Advice Rubric B, and (c) Sequence Advice Rubric B and Workplace Advice Rubric B. The sample consists of the 157 advisors.

The results are presented in Table 5. We find that advice quality seems to be more consequential for mixed-gender pairs. For example, column (1) corresponds to male advisor female advisee pairs and we estimate a typical advisor effect of 0.086 standard deviations ($p_{RI} = 0.012$). For the opposite pair, female advisor and male advisee, we estimate a typical advisor effect of 0.046 standard deviations, although we cannot rule out the null that there is no effect ($p_{RI} = 0.226$). For the other two matched-gender pairings, in column (2) and (4), we can-

not estimate advisor effects. To test if advice quality is more consequential for gender-mixed pairs we adapt the randomization inference procedure. Formally, we test the null that for female advisees the advisor gender does not matter (i.e. $\mathbf{H}_o : \sigma_\mu^f = \sigma_\mu^m$). In particular for a given advisee gender, we randomly shuffle advisor's gender, estimate new σ's, and compare them to the observed (for more details see Section B.1 of the Appendix). We report the results of this advisor gender test for both female and male advisees. We find that for female advisees, the quality of advice matters more when they have male advisors than female ($p = 0.021$). For male advisees, although directionally they have greater variability in outcomes with female advisors (0.046 standard deviations), we cannot rule out the null that there is no advisor gender difference ($p = 0.179$).

3.2 Perceptions of Advisors and Advice Performance

We explore how perceptions of advisors are related to the causal effect of their advice ($\hat{\mu}_i$). We measured advisees' perceptions of their advisor's attributes via a multi-item survey (for details see Section 1.5.1). We report the mean attribute for an advisor. Table 6 contains correlations between advisor performance ($\hat{\mu}_i$) and perceived advisor traits. We find that perceived trustworthiness correlates 0.22 with the advisor effect ($p < 0.01$), perceived competence correlates 0.16 ($p = 0.05$), and likeability correlates 0.15 ($p = 0.06$). In other words, the observed correlation for trustworthiness ($r = 0.22$) is larger than the observed correlations for the rubric measures, although we do not formally test the differences.

We asked the perception questions after advisees had done sequence task 2 and 3. Thus it is possible that the advisees who performed well rated their advisors more highly on these attributes. We run two different robustness checks.[20] First, we estimate an advisee-level leave-one-out specification. For each advisee, we calculate the advisor's mean attribute rating using only ratings from the other available advisees of that advisor. This means that if the advisee gave a rating then we drop their rating from that mean calculation. This creates a leave-one-out attribute variable for each advisee, which we add as a predictor to model (1). We find that for trustworthiness the coefficient is positively associated with post-advice performance ($\beta = 0.022$, $p < 0.01$), similarly for competence ($\beta = 0.015$, $p = 0.09$), and for likeability ($\beta = 0.016$, $p = 0.04$). This demonstrates that *other* advisees' perceptions of an advisor predict advisees' post-advice scores and not necessarily that the trait scores are driven by advisee performance. Second, we drop all advisees who answered the attribute questions (N=1974, for details see Section 1.5.1) and then re-estimate a new advisor mean residual without these observations. We still find a positive but weaker correlation coefficient for trustworthiness (Pearson's $r = 0.14$, $p = 0.08$). Similarly for for competence (Pearson's $r = 0.08$, $p = 0.34$) and likeability (Pearson's $r = 0.10$, $p = 0.23$). These robustness checks suggest that the positive correlations we find are not entirely driven by advisee post-advice performance.

[20] These are additional exploratory analyses not in our preregistration.

Table 6: Correlations of Advisor Performance and Perceived Traits

Trait	Pearson's r	p-value
Competence	0.16	0.05
Warmth	0.08	0.34
Trustworthiness	0.22	< 0.01
Dominance	-0.05	0.56
Confidence	0.01	0.89
Likeability	0.15	0.06

Notes: Entries report Pearson correlations between advisor performance ($\hat{\mu}_i$) and perceived advisor traits. The sample contains 157 advisors. All p-values are two-sided. As stated in our preregistration our aim is to explore correlations and so we do not employ a multiple-hypothesis-test correction (Parker & Weir, 2022).

4 Conclusion

In this paper we develop a novel method for identifying the causal contribution of open-ended unstructured advice. We ran a large online experiment with two types of participants, advisors and advisees. In order to study open-ended advice we designed a new task, called the sequence task, with two important features, performance is objective and there exists hidden information about how to do the task, knowledge of which can improve performance. To all advisors we disclosed this hidden information about the sequence task and then recorded their advice on how to do the task. The disclosure helps to isolate communication skills, rather than other skills such as problem solving. We repeatedly randomly assigned an advisor's audio advice to numerous advisees. We then estimate the contribution of an advisor's advice on their advisees' post-advice performance conditioning on both advisor and advisee baseline performance. We find a small but robust advisor effect. A key design feature of our approach is that the advice is delivered asynchronously and so has the potential to be scaled up to improve the precision of the individual estimates. The advice-giving skill we detect appears to be distinct from other well established individual measures such as emotion perceptiveness or fluid intelligence.

We also developed a more "naturalistic" setting in which advisors could give advice. We used a workplace vignette where the advisor had to engage in an interactive real-time conversation with an LLM-based voice chatbot. The chatbot played the role of a colleague who had recently received unexpected negative feedback at work and was seeking advice on what to do about it. As with the sequence task advice to isolate communication skills we disclosed to the advisors expert opinion about what to do with the feedback. Unlike our sequence task this setting has no objective performance metric on which to evaluate it. Instead, we use process-based qualitative evaluation rubrics to assess the advice quality. We show that an advisor's performance in the sequence task, estimated quantitatively with repeated random assignment, is not predictive of their performance in the workplace advice setting, based on a qualitative evaluation rubric. However, if we evaluate the advice given in these diverse tasks with a common rubric then we do find a correlation between the qualitative evaluation rubric scores but no correlation with

either qualitative score of the advice tasks and the advisor's quantitative score. This suggests that process-based evaluation of advice might not capture important aspects of advice that lead to improved performance.

Finally, we conduct some exploratory analysis on our quantitative advice score. First, we study the role that gender plays in our advice-giving setting, especially as it is more likely that in a given advisor-advisee pair the advisor will be male and the advisee female (Benson et al., 2026; Blau & Kahn, 2017; McKinsey & Company & LeanIn.Org, 2025; O'Brien et al., 2010). We find that advice – good and bad – is more consequential for gender-mixed pairs, in particular when the advisor is male and the advisee is female. Second, we explore associations between advisors' quantitative scores and advisees' perceptions of advisors' attributes, such as warmth and trustworthiness. We find that an advisor's trustworthiness is positively correlated with their quantitative score. This suggest that there is some predictive value of advisor performance based on advisee perceptions.

There are some limitations to our study that currently impact our results. First, the precision of our estimates. In theory the precision of our advisor effect can be arbitrarily increased, until some bound, as we already have the advisor data and can collect more advisee data. Although, in practice the financial cost is the the limiting factor. In a related point, our sample of 157 advisors is relatively small and hence there is likely measurement error in our advisor level measures, such as emotion perception, limiting the precision of our estimates. If we could reduce the cost of the advisee collection we could improve the precision of *all* our advisor-level estimates. One approach would be the leverage the data we have to train a machine-learning model to predict advisee behavior based on certain inputs such as the advice and advisee-level characteristics (Cerina & Duch, 2020). This would greatly reduce the cost of data collection, allow us collect more advisor-level data and improve the precision of our estimates.

References

Ambuehl, S., Bernheim, B. D., Ersoy, F., & Harris, D. (2025). Peer Advice on Financial Decisions: A Case of the Blind Leading the Blind? *Review of Economics and Statistics*, *107*(1), 240–255.

Bailey, P. E., Leon, T., Ebner, N. C., Moustafa, A. A., & Weidemann, G. (2023). A meta-analysis of the weight of advice in decision-making. *Current Psychology*, *42*(28), 24516–24541.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test Revised Version: A Study with Normal Adults, and Adults with Asperger Syndrome or High-functioning Autism. *Journal of Child Psychology and Psychiatry*, *42*(2), 241–251.

Benoît, J.-P., Perry, A., & Reuben, E. (2026). Performance-Feedback. *NYUAD Division of Social Science Working Paper #01010*. https://archive.nyu.edu/handle/2451/75241

Benson, A., Li, D., & Shue, K. (2026). "Potential" and the Gender Promotion Gap. *American Economic Review*, *116*(2), 375–417.

Blau, F. D., & Kahn, L. M. (2017). The Gender Wage Gap: Extent, Trends, and Explanations. *Journal of Economic Literature*, *55*(3), 789–865.

Blunden, H., Kristal, A., Whillans, A., Yoon, J., Burd, H., Bremner, G., & Yeomans, M. (2025). Eliciting Advice Instead of Feedback Improves Developmental Input. *Management Science*, *72*(3), 2247–2266.

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*(2), 127–151.

Brooks, W., Donovan, K., & Johnson, T. R. (2018). Mentors or Teachers? Microenterprise Training in Kenya. *American Economic Journal: Applied Economics*, *10*(4), 196–221.

Bucher-Koenen, T., Hackethal, A., Koenen, J., & Laudenbach, C. (2025). Gender Differences in Financial Advice. *American Economic Review*, *115*(12), 4218–4252.

Burleson, B. R. (1982). The Development of Comforting Communication Skills in Childhood and Adolescence. *Child Development*, *53*(6), 1578–1588.

Casaburi, L., & Heursen, L. (2026). The Two Sides of Peer Feedback: Strategic Provision and Selective Use.

Cerina, R., & Duch, R. (2020). Measuring public opinion via digital footprints. *International Journal of Forecasting*, *36*(3), 987–1002.

Chatterji, A., Delecourt, S., Hasan, S., & Koning, R. (2019). When does advice impact startup performance? *Strategic Management Journal*, *40*(3), 331–356.

Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88–97.

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational Judgment Tests: Constructs Assessed and a Meta-Analysis of their Criterion-Related Validities. *Personnel Psychology*, *63*(1), 83–117.

Cullen, Z., & Perez-Truglia, R. (2023). The Old Boys' Club: Schmoozing and the Gender Gap. *American Economic Review*, *113*(7), 1703–1740.

Dimitriadis, S., & Koning, R. (2022). Social Skills Improve Business Performance: Evidence from a Randomized Control Trial with Entrepreneurs in Togo. *Management Science*, *68*(12), 8635–8657.

Dweck, C. S., & Leggett, E. L. (1988). A Social-Cognitive Approach to Motivation and Personality. *Psychological Review*, *95*(2), 256–273.

Emanuel, N., Harrington, E., & Pallais, A. (2026). The Power of Proximity to Coworkers. *The Quarterly Journal of Economics*, *141*(3), 1825–1870.

Ernst, M. D. (2004). Permutation Methods: A Basis for Exact Inference. *Statistical Science*, *19*(4).

Eskreis-Winkler, L., & Fishbach, A. (2020). Hidden failures. *Organizational Behavior and Human Decision Processes*, *157*, 57–67.

Eskreis-Winkler, L., Fishbach, A., & Duckworth, A. L. (2018). Dear Abby: Should I Give Advice or Receive It? *Psychological Science*, *29*(11), 1797–1806.

Feng, B., & MacGeorge, E. L. (2010). The Influences of Message and Source Factors on Advice Outcomes. *Communication Research*, *37*(4), 553–575.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83.

Gnepp, J., Klayman, J., Williamson, I. O., & Barlas, S. (2020). The future of feedback: Motivating performance improvement through future-focused feedback (P. Iannello, Ed.). *PLOS ONE*, *15*(6), e0234444.

Graeber, T., Noy, S., & Roth, C. (2026). The Transmission of Reliable and Unreliable Information. *Working Paper*.

Green, A. (2021, February). Mentoring a very timid employee, secretary asks every day if I've been vaccinated yet, and more. https://www.askamanager.org/2021/04/mentoring-a-very-timid-employee-secretary-asks-every-day-if-ive-been-vaccinated-yet-and-more.html

Heslin, P. A., Vandewalle, D., & Latham, G. P. (2006). Keen to Help? Managers' Implicit Person Theories and their Subsequent Employee Coaching. *Personnel Psychology*, *59*(4), 871–902.

Heursen, L., Friess, S., & Chugunova, M. (2026). Reputational Concerns and Advice-Seeking at Work.

High, A. C., & Dillard, J. P. (2012). A Review and Meta-Analysis of Person-Centered Messages and Social Support Outcomes. *Communication Studies*, *63*(1), 99–118.

Hoffman, M., & Tadelis, S. (2021). People Management Skills, Employee Attrition, and Manager Rewards: An Empirical Analysis. *Journal of Political Economy*, *129*(1), 243–285.

Hospido, L., Laeven, L., & Lamo, A. (2022). The Gender Promotion Gap: Evidence from Central Banking. *The Review of Economics and Statistics*, *104*(5), 981–996.

Hüning, H., Mechtenberg, L., & Wang, S. W. (2022). Using Arguments to Persuade: Experimental Evidence. *Working Paper*.

Hyland, M., Djankov, S., & Goldberg, P. K. (2020). Gendered Laws and Women in the Workforce. *American Economic Review: Insights*, *2*(4), 475–490.

Jones, N., Marks, R., Ramirez, R., & Rios-Vargas, M. (2021). 2020 Census Illuminates Racial and Ethnic Composition of the Country. https://www.census.gov/library/stories/2021/08/improved-race-ethnicity-measures-reveal-united-states-population-much-more-multiracial.html

Kleven, H., & Landais, C. (2017). Gender Inequality and Economic Development: Fertility, Education and Norms. *Economica*, *84*(334), 180–209.

Lievens, F., & Sackett, P. R. (2012). The validity of interpersonal skills assessment via situational judgment tests for predicting academic success and job performance. *Journal of Applied Psychology*, *97*(2), 460–468.

MacGeorge, E. L., Feng, B., Butler, G. L., & Budarz, S. K. (2004). Understanding Advice in Supportive Interactions.: Beyond the Facework and Message Evaluation Paradigm. *Human Communication Research*, *30*(1), 42–70.

MacGeorge, E. L., Guntzviller, L. M., Hanasono, L. K., & Feng, B. (2016). Testing Advice Response Theory in Interactions With Friends. *Communication Research*, *43*(2), 211–231.

MacGeorge, E. L., & Van Swol, L. M. (2018a, May). Advice Across Disciplines and Contexts. In E. L. MacGeorge & L. M. Van Swol (Eds.), *The Oxford Handbook of Advice* (1st ed., pp. 3–18). Oxford University Press.

MacGeorge, E. L., & Van Swol, L. M. (Eds.). (2018b, July). *The Oxford Handbook of Advice* (1st ed.). Oxford University Press.

McAleer, P., Todorov, A., & Belin, P. (2014). How Do You Say 'Hello'? Personality Impressions from Brief Novel Voices (C. R. Larson, Ed.). *PLoS ONE*, *9*(3), e90779.

McKinsey & Company & LeanIn.Org. (2025). *Women in the Workplace*. https://womenintheworkplace.com

Meshi, D., Biele, G., Korn, C. W., & Heekeren, H. R. (2012). How Expert Advice Influences Decision Making (J. Daunizeau, Ed.). *PLoS ONE*, *7*(11), e49748.

Minni, V. (2026). Making the Invisible Hand Visible: Managers and the Allocation of Workers to Jobs. *The Quarterly Journal of Economics*, *141*(3), 1871–1920.

Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. *Political Analysis*, *16*(4), 372–403.

Ng, Y. X., Koh, Z. Y. K., Yap, H. W., Tay, K. T., Tan, X. H., Ong, Y. T., Tan, L. H. E., Chin, A. M. C., Toh, Y. P., Shivananda, S., Compton, S., Mason, S., Kanesvaran, R., & Krishna, L. (2020). Assessing mentoring: A scoping review of mentoring assessment tools in internal medicine between 1990 and 2019 (A. M. Samy, Ed.). *PLOS ONE*, *15*(5), e0232511.

O'Brien, K. E., Biga, A., Kessler, S. R., & Allen, T. D. (2010). A Meta-Analytic Investigation of Gender Differences in Mentoring. *Journal of Management*, *36*(2), 537–554.

OECD. (2025, September). *Gender Equality in a Changing World: Taking Stock and Moving Forward*.

Oswald, W., & Roth, E. (1978). *Der Zahlen-Verbindungs-Test*. Hogrefe.

Parker, R. A., & Weir, C. J. (2022). Multiple secondary outcome analyses: Precise interpretation is important. *Trials*, *23*(1), 27.

Pulford, B. D., Colman, A. M., Buabang, E. K., & Krockow, E. M. (2018). The persuasive power of knowledge: Testing the confidence heuristic. *Journal of Experimental Psychology: General*, *147*(10), 1431–1444.

Raven, J., & Raven, J. (2003). Raven Progressive Matrices. In R. S. McCallum (Ed.), *Handbook of Nonverbal Assessment* (pp. 223–237). Springer US.

Sandvik, J., Saouma, R., Seegert, N., & Stanton, C. (2020). Workplace Knowledge Flows. *The Quarterly Journal of Economics*, *135*(3), 1635–1680.

Sandvik, J., Saouma, R., Seegert, N., & Stanton, C. (2026). Should Human Capital Development Programs be Mandatory or Voluntary? Evidence from a Field Experiment on Mentorship. *Management Science*, *72*(6), 4610–4632.

Schotter, A. (2023, March). *Advice, Social Learning and the Evolution of Conventions* (1st ed.). Cambridge University Press.

Serra-Garcia, M. (2026). The Attention-Information Trade-Off. *American Economic Review*, *116*(5), 1579–1610.

Shen, W., Liu, Y. (, & Wang, Y. ((2025). Penalty or Reward? The Effect of Social Disincentives on Online Users' Contributions. *Management Science*, *71*(8), 6770–6792.

Sniezek, J. A., & Buckley, T. (1995). Cueing and Cognitive Conflict in Judge-Advisor Decision Making. *Organizational Behavior and Human Decision Processes*, *62*(2), 159–174.

Thaler, M., Toma, M., & Wang, V. Y. (2025). Numbers Tell, Words Sell. *Working Paper*.

Van Swol, L. M. (2011). Forecasting another's enjoyment versus giving the right answer: Trust, shared values, task effects, and confidence in improving the acceptance of advice. *International Journal of Forecasting*, *27*(1), 103–120.

Van Swol, L. M., Paik, J. E., & Prahl, A. (2018, May). Advice Recipients: The Psychology of Advice Utilization. In E. L. MacGeorge & L. M. Van Swol (Eds.), *The Oxford Handbook of Advice* (1st ed., pp. 21–42). Oxford University Press.

VandeWalle, D., Cron, W. L., & Slocum, J. W. (2001). The Role of Goal Orientation Following Performance Feedback. *Journal of Applied Psychology*, *86*(4), 629–640.

Vernon, P. A. (1993). Der Zahlen-Verbindungs-Test and other trail-making correlates of general intelligence. *Personality and Individual Differences*, *14*(1), 35–40.

Vu, N. C., Li, Y., & High, A. C. (2025). Advice Response Theory: A Meta-Analytic Review. *Communication Research*, 00936502251351041.

Weidmann, B., & Deming, D. J. (2021). Team Players: How Social Skills Improve Team Performance. *Econometrica*, *89*(6), 2637–2657.

Weidmann, B., Vecci, J., Said, F., Bhalotra, S., Adhvaryu, A., Nyshadham, A., Tamayo, J., & Deming, D. (2026). How Do You Identify a Good Manager? *The Quarterly Journal of Economics*, *141*(2), 1581–1633.

Weidmann, B., & Xu, Y. (2025). Measuring Emotion Perception Ability Using AI-Generated Stimuli: Development and Validation of the PAGE Test. *Journal of Intelligence*, *13*(9), 116.

Weidmann, B., Xu, Y., & Deming, D. J. (2025). Measuring Human Leadership Skills With AI Agents. *NBER Working Paper 33662*. http://www.nber.org/papers/w33662

Yaniv, I., & Choshen-Hillel, S. (2012). When guessing what another person would say is better than giving your own opinion: Using perspective-taking to improve advice-taking. *Journal of Experimental Social Psychology*, *48*(5), 1022–1028.

Yaniv, I., & Kleinberger, E. (2000). Advice Taking in Decision Making: Egocentric Discounting and Reputation Formation. *Organizational Behavior and Human Decision Processes*, *83*(2), 260–281.

Yin, J., Ngiam, K. Y., Tan, S. S.-L., & Teo, H. H. (2025). Designing AI-Based Work Processes: How the Timing of AI Advice Affects Diagnostic Decision Making. *Management Science*, *71*(11), 9361–9383.

Appendix

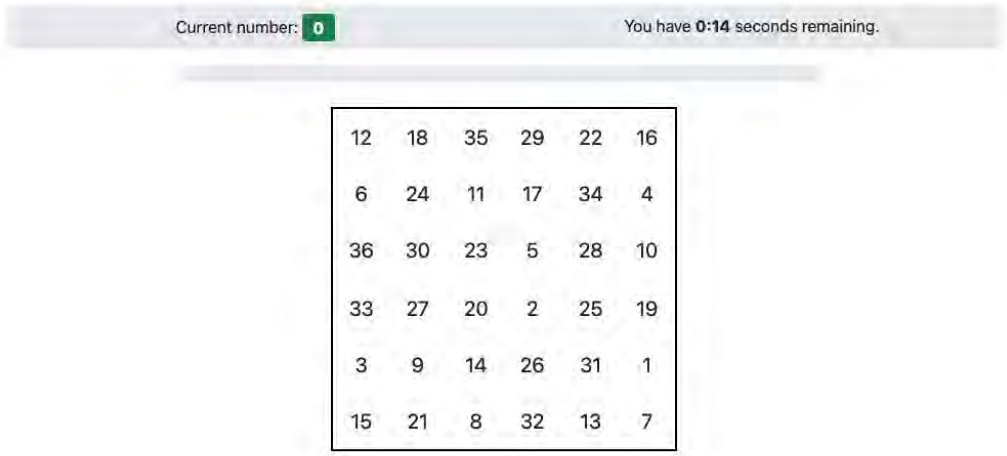
A Experiment Materials

A.1 The Sequence Task and Advice Recording

The sequence task is an adapted version of the Der Zahlen-Verbindungs-Test (ZVT) or the number connection task (Oswald & Roth, 1978). The original ZVT consisted of numbers 1 to 90 placed in a square randomly, the participant then has to draw a line connecting all the numbers sequentially as quickly as possible; their score being the amount of time is takes them to accomplish this. There are a number of variations on this, such as going in a decreasing order or using letters of the alphabet (Vernon, 1993). This task was developed to measure an information-processing model of intelligence, which views speed of information-processing and the duration of short-term memory as key factors to understanding the cause of individual differences in intelligence.

Both advisors and advisees did three computer based sequences tasks, where they clicked on each number in an increasing order starting at 1. A valid click had to be close enough to the number, this was indicated by a green ring appearing around the number. Figure A1 is a screen shot of a participants screen. Participants were provided with the following information: Their current number in the sequence, their progress indicated by the bar above the sequence task, and their time remaining. Only if they clicked on the next number in the sequence would the current number update, otherwise it remained the same. They did sequence tasks of either 24 or 36 digits and had 20 seconds to do the task

Figure A1: Sequence task experimental interface

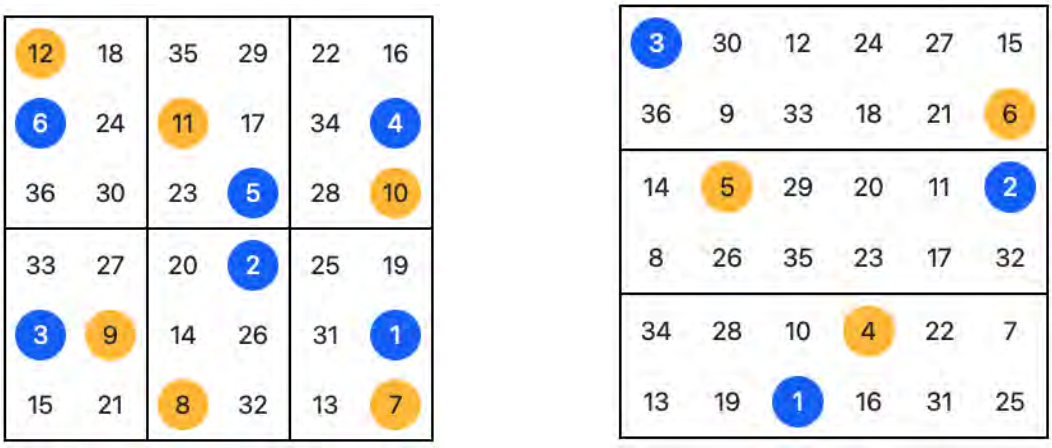


Notes: Figure shows a sequence task of length 36 with the interface that participants saw. Current number tells the participant what number in the sequence they have got to with the bar, above the sequence task, giving a visual indication of their progress. They also saw their time remaining for the task.

After advisors finished the first sequence task they were all provided with identical information about the underlying grid structure and placement pattern of the first sequence task they had done. They were told that identifying the underlying structure and placement pattern made

it easier to complete the task. They were shown the hidden grid structure of first sequence task (Figure A2a) as well as other example sequence tasks to demonstrate that there were numerous possible grid structures and placement patterns (Figure A2b).

Figure A2: The sequence tasks of the study



(a) (b)

Notes: Panel (a) Advisors were shown this image after they did the first sequence task, showing them the hidden grid structure and placement pattern. Panel (b) is an example of another sequence task with different grid structure and placement pattern, this image as well as others were shown to advisors before they attempted the second sequence task. Only advisors saw this information.

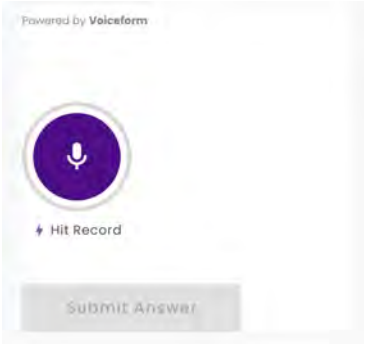
Subsequent to the disclosure about the hidden structure, advisors completed the second and third sequence tasks. Immediately after completing the third sequence task they made their audio recording of advice on how to do the sequence task. We showed them instructions emphasizing that their advice would be shared with a randomly selected group of advisees, who would not have seen the same hidden information that had been disclosed to them and that the incentives of the advisor and advisee were aligned, both types of participants bonus could depend on the advisees performance. Finally we reiterated that advisors should not disclose personal information or use disrespectful language as that risked them being disqualified from the study. They were able to type notes on what they would say. Figure A3 shows the audio recording software interface.

The advisees did the same three sequence tasks as advisors in the same order, except after the first sequence task they did not see the information shared with the advisors. Instead they were able to listen to an audio recording of advice from an advisor, before then proceeding onto the second and third sequence tasks. We informed the advisees before they listened to the advice that their advisor had been given access to special information about the sequence task and that the advisor's bonus could depend on their performance.

A.2 Workplace Advice Vignette and Conversation Recording

As our sequence task advice setting is abstract, we wanted to measure advice-giving skills in a more naturalistic but still economically relevant setting i.e. a workplace dilemma. To do this we created a vignette of a workplace dilemma where participants (Advisors) read a vignette

Figure A3: Sequence Advice Recording Software



Notes: The Voiceform software to make an audio recording. Participants pressed the microphone sign to start and stop the recording. Once they were done they could press "Submit Answer" to finish the recording.

(see Figure A4a) as well as a set of expert recommendations (see Figure A4b). This is followed by a 5-minute voice conversation with an LLM-based chatbot, who plays the protagonist of the vignette, Casey (see Figures A4c and A4d). The participant's role is as an advisor/mentor to Casey. By giving the advisors the expert recommendations we mimic the quantitative advice measure and isolate advice-giving from other skills such as idea generation.

A.3 Measure of Individual Skills and Characteristics

We measured a number of individual skills and characteristics of the advisors: Emotional perception, fluid intelligence, mindsets, and social impressions. This section provides details on how we collected data on each of these.

Advisor's emotion perceptiveness was measured using PAGE (perceiving AI generated emotions test, see Weidmann & Xu, 2025). An example problem is provided in Figure A5a. The short version, with 16 faces, is used in this study. Participants must select which word best describes the emotion in the picture. There are 6 choices and they can click on the ? icon to see definitions of the emotions if they are unsure. Advisor's fluid intelligence was measured using a set of Ravens Advanced Progressive Matrices (Raven & Raven, 2003). An example item is provided in Figure A5b. In the example participants are asked to examine the patterns established by the rows and columns of eight images, and to indicate which of the options, labeled 1 to 8, best fits the pattern, and fills the vacant item space.

We measured advisors' beliefs about the malleability of other peoples' intelligence and personality (for example see, Dweck & Leggett, 1988). Advisors answered two items both made up of 3 sub-items: one related to intelligence (i-iii) and one related to personality (iv-vi). All sub-item responses are on a Likert scale with the following options: Strongly Agree, Agree, Neither Agree nor Disagree, Disagree, Strongly Disagree. The order of all the 6 sub-items was randomized to mitigate order effects.

- Intelligence

  i Intelligence is something very basic about a person and it can't be altered much.

  ii People can learn new things, but they can't really change how intelligent they are.

Figure A4: Workplace Vignette



(a) (b)

(c) (d)

Notes: Panel (a) contains the background details to the workplace scenario, outlining the advisors role as Casey's mentor. Panel (b) shows the expert assessment of Casey's feedback and that this is the information they need to communicate to Casey. They could also write notes to use during their conversation. Panels (c) and (d) show the LLM conversation interface with Casey, where to aid the flow of conversation the circle changes color to indicate the different modes of Casey e.g. green for listening when the participant is speaking.

iii Everyone has a certain kind of intelligence and there is not much that can be done to really change that.

- Personality

iv The kind of person someone is, is something very basic about them, and it can't be altered much.

v People can do things differently, but the important parts of who they are can't really be changed.

vi Everyone is a certain kind of person and there is not much that can be done to really change that.

Figure A5: Example items from measures of individual skill



(a) (b)

Notes: Panel (a) shows a single question from the emotion perception task PAGE (Weidmann & Xu, 2025). Panel (b) shows a single question from the Ravens Matrix (Raven & Raven, 2003).

Finally, we measure advisees' perception of their advisor, since we have 64 advisees per advisor we randomly sample a subset of the advisees. We measure the following six attributes: warmth, competence, trustworthiness, dominance, likeablity, and confidence (Fiske et al., 2007; McAleer et al., 2014). For each attribute X we asked "How X do you think they are?". We used a Likert scale from not at all to extremely (Not at all, Slightly, Moderately, Very, Extremely). The order of the questions was randomized.

B Additional Details

B.1 Inference

In our main analyses we estimate the magnitude of the typical advisor effect ($\hat{\sigma}_\mu$) using equation (3). We compare this to the null hypothesis that advisors have no impact on their advisees after controlling for the advisor-advisee pair of baseline sequence task skills. Our preregistered inferential approach is to calculate p-values using randomization inference. For robustness we also report alternative estimates of uncertainty using a Wald estimator and Profile-likelihood confidence intervals.

The randomization inference procedure has four steps. First, we estimate model (1) and then model (3) giving our observed $\hat{\sigma}_\mu$. Second, we simulate 5,000 reassignments of advisees that preserve the original blocked and balanced assignment procedure of advisees, so that an advisor has the same composition of advisees as they did in the experiment. For each reassignment, we re-estimate the random-effects model (3) and compare the resulting estimate of $\hat{\sigma}_{\mu(NULL)}$ with the observed estimate. The one-sided p-value is the proportion of simulated estimates greater than the observed estimate (Ernst, 2004).[A1]

[A1] In the preregistration we follow the inference procedure of Weidmann and Deming (2021). However, we made two accidental errors in the text. First, we incorrectly said that we would re-estimate both model (1) and model (3) for each permutation. The actual procedure is to only re-estimate model (3) for each permutation. Second, we incorrectly stated that our one-sided p-value is the proportion of simulated estimates *at least as large as* the observed estimate. When this should've been be a strict inequality i.e., *greater than*.

We adapt the randomization inference procedure to determine if estimates of the advisor effect for advisees vary by advisor gender. We do this separately for female and male advisees. The procedure for female advisees is the following (the procedure for male advisees follows the same steps except the first). First, we drop all male advisee observations. Second, we split the data into females advisees with female advisors and those with male advisors. For both groups we estimate the model (1) and then model (3). We take the difference of the observed $\hat{\sigma}$'s and call that $\Delta\hat{\sigma}_{\mu}$. Third, we simulate 5,000 reassignments of the advisor gender, note this means they have the same advisees. For each reassignment, we re-estimate the random-effects model (3) for both advisor genders, compute $\Delta\hat{\sigma}_{\mu(NULL)}$, and then compare this estimate with the observed estimate. The one-sided p-value is the proportion of simulated estimates greater than the observed estimate.

B.2 Completion Criteria

We preregistered that we would recruit 157 advisors and 64 advisees per advisor (10,048 advisees in total). Our preregistration also included an exclusion criteria to ensure the advisor and advisee submissions were valid (see Table B1). In addition we had an audio check and understanding questions that both advisors and advisees needed to pass in order to proceed with the study. 241 advisors consented to the study, but 47 failed the understanding questions. Of the 194 who passed the understanding 184 completed the study, but 27 were not used as they were invalid based on the criteria of Table B1. For advisees of the 10,919 who consented to the study, 752 either failed the understanding questions or returned their submission. Of the 10,167 who passed the understanding 119 failed to complete all three sequence tasks. Advisors and advisees whose submissions did not meet the criteria were excluded from the analysis.

Before being presented to advisees, valid audio recordings were trimmed to remove leading and trailing silence, processed with a noise gate of $-40\,\text{dB}$ to reduce background noise, and exported at 256 kbps. Recordings were not otherwise edited in a way that would have changed their substantive content.

B.3 Text Analysis

To explore how the language advisors used varied by advisor performance we categorize the advisors based on their estimated advisor effect from equation (2) into two groups, high and low. High corresponds to the top 20% of the distribution and low the bottom 20%, with rounding this gives 32 advisors in each group. Then we compare the frequency of the words across these two groups based on an approach to feature extraction outlined in Monroe et al. (2008). We use all 157 advisor's sequence task advice to estimate the corpus distribution of word usage.

We conducted pre-processing for this procedure. First we as the sequence task is a very specific task we manually preserved certain words: for example task-relevant spatial terms "left", "right" "middle", "top", and "bottom" and numbers written as words were normalized to numerical digits. Second, we tonkenized, removed stop words and lemmatized the tokens, e.g. "pattern" and "patterns" count as one token.

We identified words characteristic of the two performance groups using the log-odds method with an informative Dirichlet prior Monroe et al. (2008). Let y_{wg} denote the number of oc-

Table B1: Data-quality and exclusion criteria

Sample	**Criterion**	**Consequence**
Advisor	The advice recordings contains no intelligible English speech.	The quantitative-advice recording will not be assigned to advisees, and the advisor submission will be excluded.
Advisor	The advice recordings contains personally identifying information, offensive language, or other content that cannot be shared with Prolific participants.	The quantitative-advice recording will not be assigned to advisees, and the advisor submission will be excluded.
Advisor	The participant does not complete the required components: the three sequence tasks, quantitative-advice task, qualitative-advice task, emotion-perception task, IQ task, study questions, background questions.	The quantitative-advice recording will not be assigned to advisees, and the advisor submission will be excluded.
Advisor or advisee	The participant fails the honeypot check or provides clear evidence of automated or non-engaged participation, such as making no attempt to complete the sequence task.	The submission will be excluded from the analysis; excluded quantitative-advice recordings will not be assigned to advisees.

currences of lemma w in performance group g, and let N_g denote the total number of retained tokens in that group. We constructed the informative prior from the complete corpus of 157 advice documents, setting α_w equal to the full-corpus frequency of lemma w and $\alpha_0 = \sum_w \alpha_w$. The smoothed log odds for lemma w in group g were calculated as

$$\ell_{wg} = \log\left[\frac{y_{wg} + \alpha_w}{N_g + \alpha_0 - y_{wg} - \alpha_w}\right].$$

We then calculated the difference between the top- and bottom-group log odds,

$$\delta_w = \ell_{wT} - \ell_{wB},$$

and standardized this difference using the approximate variance

$$V_w = \frac{1}{y_{wT} + \alpha_w} + \frac{1}{y_{wB} + \alpha_w}.$$

The resulting weighted log-odds statistic was

$$z_w = \frac{\delta_w}{\sqrt{V_w}}.$$

Positive values indicate that a lemma was relatively more characteristic of high-performing advisors, whereas negative values indicate that it was relatively more characteristic of low-performing advisors. Larger absolute values indicate a stronger difference. To limit the influence

of idiosyncratic vocabulary, we retained only lemmas appearing in at least five documents across the high and low advisor groups. Figure 4 contains the 20 lemmas with the largest absolute log-odds scores in both the two groups.

B.4 Advice Evaluation Rubrics

We have three different evaluation rubrics A, B, and C.[A2] For all the evaluation rubrics we use the same procedure to evaluate the advice text of advisors. The only difference between the rubrics are the details of the scoring criteria contained in the prompts. In general, to score each advisor's text we use a large language model acting as a judge. The model used was gpt-5.6-sol at medium reasoning effort, called through the OpenAI Responses API with a fixed prompt for each rubric and a JSON output schema. Specifically, two independent raters score every conversation. Each rater is a separate API call with an identical prompt, and neither sees the other's scores or reasoning. Where the two raters differ by two or more points on any single dimension, we call a third rater, which is also blind to the first two, and resolve each dimension by the median of the available ratings. Furthermore, where appropriate we provide the LLM statistics about the advice e.g. word length or number of conversational turns.

In the next sections we give the details of each of the prompts.

B.4.1 Prompt A

This evaluation rubric is based on our preregistration and only evaluates the workplace advice. We evaluated the advice text on the following four dimensions: 'Acceptance', 'Future', 'Sustainability' and 'Responsiveness'.[A3] See the exact prompt at the end of this section.

We scored the four dimensions on a 0–5 integer scale. 'Acceptance' asks whether the advisor helps Casey engage seriously with the manager's feedback without demanding either automatic obedience or reflexive dismissal. 'Future' asks whether the advisor gives Casey concrete, usable actions for the next few weeks, and folds in the specificity of the examples offered. 'Sustainable' asks whether the advice finds a workable middle ground between greater visibility and Casey's careful, prepared working style. 'Responsiveness' asks whether the advisor engages with what Casey actually says.

Adjudication was used once out of 157 total evaluations. Inter-rater agreement is high. Exact agreement between the two raters is 79% on acceptance, 89% on future, 83% on sustainable, and 78% on responsiveness, with quadratic-weighted kappas of 0.85 on acceptance, 0.90 on future, 0.93 on sustainable, and 0.82 on responsiveness. The two-rater mean has a Spearman–Brown reliability of 0.97.

```
    <YOUR ROLE>
You are an expert evaluator of advice-giving conversations. Judge the
   practical quality of the ADVISOR's advice. We care about whether the
    advice would genuinely help Casey, not whether it sounds polished
   or follows a textbook script.
```

[A2] In our preregistration we included rubric A, rubric B and C are additional exploratory rubrics.

[A3] This is slight deviation from our preregistration, which has Warmth instead of Responsiveness, but given the back and forth nature of the interaction we feel this is a better performance dimension to measure as it captures the advisor's engagement with the LLM chatbot rather than just repeating stock phrases.

```
<SCENARIO THE ADVISOR IS RESPONDING TO>
Casey's work was rated positively, but Casey's manager (Alex) said
   Casey needs to be more proactive and visible - speaking up in
   meetings, sharing work in progress, and taking initiative.
Casey was surprised and sees themself as a careful, thorough worker who
    prefers preparation and polished work. A follow-up review is in two
    weeks.

You will receive a transcript with numbered turns labelled ADVISOR: and
    CASEY:, plus exact interaction statistics.

This is spoken conversation transcribed automatically. Expect
   disfluencies and false starts. Judge substance, not fluency or
   polish.

You are scoring the ADVISOR throughout.

<GLOBAL CALIBRATION>
Score four equally important dimensions: acceptance, future,
   sustainable, and responsiveness.
The examples named in the scenario (sharing drafts, sending updates,
   speaking in meetings) are context, NOT an answer key.
Do not reward an advisor merely for repeating these ideas. Judge
   whether the advice is genuinely useful and appropriately fitted to
   Casey.
A short, informal answer can be excellent. A long, polished answer can
   be poor.
Do not reward length, repetition, confident language, or managerial-
   sounding phrasing.
Do not assume either Casey or Alex is necessarily right. Good advice
   can take both perspectives seriously while helping Casey decide what
    to do.
Score each criterion independently.

Use this 0-5 scale for every criterion. Treat 3 as a real quality
   threshold, not a default:
0 = actively counterproductive on this criterion, or completely fails
   it.
1 = largely poor: mostly irrelevant, generic, ill-judged, or undermined
    elsewhere in the conversation.
2 = mixed / limited: relevant and somewhat useful, but too generic,
   incomplete, or poorly calibrated.
3 = solid: materially useful advice that fits Casey's actual situation.
4 = strong: specific, well calibrated, and clearly tailored to Casey's
   constraints.
5 = exceptional: consistently strong, precise, adaptive, and
   practically useful across the conversation. Use 5 rarely.
```

```
<CRITERIA>

acceptance - BALANCED HANDLING OF THE FEEDBACK.
Judge whether the advisor helps Casey engage seriously with the
   feedback without demanding automatic obedience or reflexive
   dismissal.
High scores preserve Casey's agency: extract a potentially useful
   signal, distinguish the manager's observation from the manager's
   interpretation, and/or suggest clarifying what 'proactive' or '
   visible' should mean in practice.
It is fine to validate that the feedback may be unfair or subjective,
   but the conversation should move toward deciding what Casey can
   usefully do.
Lower scores for telling Casey simply to ignore the feedback, or at the
    other extreme for insisting Casey must accept it as true, be
   grateful for it, stop feeling upset, or trust the manager's motives
   without evidence.
A generic 'keep an open mind' or 'take feedback to heart' with no
   useful reasoning is at most a 2.

future - CONCRETE, USEFUL FORWARD ACTION.
Judge whether the advisor gives Casey practical actions they could
   realistically take from now on, especially over the next few weeks.
This criterion includes both the quality of the plan and the
   concreteness of the examples.
High scores give Casey at least one well-chosen, usable action and
   enough detail to understand how to carry it out.
Useful detail might include what Casey should do, when to do it, what
   to say or include, what situation should trigger the action, or how
   to tell whether it is working.
One genuinely useful and well-developed action can score highly. A long
    list is not required.
Simply saying 'share drafts', 'send updates', 'speak up more', or 'take
    more initiative' is not enough for a high score unless the advisor
   makes the suggestion useful in Casey's particular situation.
Similarly, vague advice such as 'focus on the future' or 'just be more
   confident' is at most a 2.
When deciding between 2 and 3, ask: after this conversation, does Casey
    genuinely know something useful they could do differently? If not,
   score 2 or below.

sustainable - FIT WITH CASEY'S WORKING STYLE AND REAL CONSTRAINTS.
Judge whether the advice finds a workable middle ground between greater
    visibility and Casey's careful, prepared style.
High scores make visibility easier to sustain: use preparation, limited
    frequency, relevant moments, small steps, or other methods that let
    Casey remain thoughtful rather than perform constantly.
```

```
Do not require Casey to become more extroverted, spontaneous,
   relentlessly confident, or willing to share everything unfinished.
It is acceptable to acknowledge that workplace visibility can be partly
    performative, as long as the proposed response is realistic and
   does not ask Casey to misrepresent facts or abandon good work
   practices.
Generic reassurance such as 'just be yourself' or generic pressure such
    as 'just speak up' is at most a 2 unless translated into a workable
    approach.

RESPONSIVENESS - two steps.

Step 1, turn-level uptake. The prompt lists every ADVISOR turn that
   directly follows a CASEY turn.
Return exactly one entry for each listed turn number.
Code whether that advisor turn responds to the substance of Casey's
   immediately preceding turn:
  0 = misses an important unresolved point: ignores the main concern,
      answers a different question, continues a prior script without
      engaging a substantively new issue, or leaves an important
      question unanswered.
  1 = partial uptake or optional elaboration: addresses the topic only
      partly; OR Casey asks for further examples, scripts, wording, or
      refinement after the advisor has already given a reasonably usable
       answer on the same issue.
  2 = direct uptake: answers an important unresolved question,
      addresses a substantively new concern or objection, uses
      information Casey just supplied, or meaningfully adapts the advice
       because of Casey's turn.

DIMINISHING OBLIGATION FOR REPEATED REQUESTS:
Casey may repeatedly ask for more specificity, especially later in the
   conversation.
Once the advisor has already given a reasonably usable answer on a
   topic, later requests for another example, script, exact phrase,
   wording variation, or refinement are optional elaboration rather
   than a new unmet need.
Failing to provide more detail in those cases should usually receive
   uptake = 1 rather than 0 and should have little effect on the
   overall responsiveness score.
This is especially important near the end of the conversation. The
   advisor does not need to continue generating examples indefinitely.
However, this does NOT apply when Casey raises a genuinely new concern,
    identifies a problem with the earlier advice, reveals a new
   constraint, says the earlier advice is not workable, or asks
   something necessary to make the advice usable.
Those substantive contributions should be judged normally even if they
   occur at the end.
```

```
Step 2, assign the overall 0-5 responsiveness score.
Base this mainly on engagement with Casey's important and substantively
    new contributions:
  0 = essentially non-responsive: the advice proceeds regardless of
     what Casey says.
  1 = mostly scripted or repetitive: Casey's substantive contributions
     rarely affect the response.
  2 = mixed: some direct engagement, but important unresolved concerns
     are missed or answered generically.
  3 = solid: most important or substantively new Casey turns receive a
     relevant response.
  4 = strong: Casey's concerns, constraints, or objections repeatedly
     and visibly shape the advice.
  5 = exceptional and sustained: the advisor consistently follows and
     adapts to Casey, with no important unresolved issue left hanging.

Weight important unresolved concerns much more heavily than repeated
   requests for extra specificity.
A conversation can receive a 4 or 5 even if some late-stage requests
   for optional elaboration are not answered.

Interaction statistics are SECONDARY evidence only. Do not mechanically
    reward a particular airtime split or punish a short conversation.

<FINAL SCORING DISCIPLINE>
Before assigning each score, consider both the strongest evidence for
   and against a high score.
Do not give 4-5 merely because the advisor mentions the right topic.
Do not use Casey's positive-sounding reactions as proof that the advice
    was good.
Do not infer advice the advisor never gave.
Judge whether Casey received enough useful guidance to act, not whether
    the advisor exhausted every possible follow-up request.

For each of the four criteria, first quote the single most decisive
   stretch of ADVISOR speech
(verbatim, at most 40 words; use "NONE" if absent), then give a one-
   sentence rationale, then
the integer score.
```

B.4.2 Prompt B

This evaluation rubric was not included in our preregistration but was developed to allow for a "common" evaluation of the advice text to explore the relationship between our advisor performance metric and subjective evaluation metrics of the two advice settings. The rubric was generated by giving an LLM the advice text from both settings and asking it to come up with

common criteria, the LLM did not see the advisor performance score. As this was a designed to be a common criteria we gave the LLM the information about the different contexts of the two advice settings to allow it to be incorporated into the evaluation.

We scored the advice text on the following five dimensions: 'Relevance', 'Actionability', 'Rationale', 'Feasibility', and 'Clarity'. See the exact prompt at the end of this section. We score five dimensions, each on a 0–4 integer scale. 'Relevance' asks whether the advice fits the recipient's goal. 'Actionability' asks whether the advice gives the recipient concrete steps. 'Rationale' asks whether the advice explains why it should be followed. 'Feasibility' asks whether the advice is realistic and takes into account the settings constraints. 'Clarity' asks whether the advice is understandable and internally consistent.

Adjudication was used three times out of 314 total evaluations.[A4] In the sequence task advice inter-rater agreement is high. Exact agreement between the two raters is 77% on relevance, 82% on actionability, 80% on rationale, 75% on feasibility, and 89% on clarity, with quadratic-weighted kappas of 0.77 on relevance, 0.84 on actionability, 0.83 on rationale, 0.64 on feasibility, and 0.79 on clarity. The two-rater mean has a Spearman–Brown reliability of 0.94. In the workplace advice inter-rater agreement is high. Exact agreement between the two raters is 83% on relevance, 90% on actionability, 87% on rationale, 85% on feasibility, and 89% on clarity, with quadratic-weighted kappas of 0.76 on relevance, 0.85 on actionability, 0.87 on rationale, 0.87 on feasibility, and 0.75 on clarity. The two-rater mean has a Spearman–Brown reliability of 0.96.

```
LLM PROMPT: COMMON RUBRIC FOR EVALUATING ADVICE

You are evaluating the text-based quality and likely effectiveness of
   advice in
two different contexts:

1. Quantitative-task advice: advice about completing a specific timed
   number-
   sequence task. Each row in quant_advice_transcript.csv contains one
      advice
   text in advice_transcript.

2. Casey workplace advice: advice about responding to workplace
   feedback,
   becoming more proactive and visible, and doing so without abandoning
       a careful
   and thorough work style. In casey_transcripts.csv, "user" is the
      advice giver;
   "assistant_male" and "assistant_female" are the advice recipient.
      Evaluate all
   user turns for one prolific_pid together as one advice episode. Use
      assistant
   turns only to understand the recipient's situation, concerns, and
      requests.
```

[A4] Each advisor has two evaluations here, one for each advice setting.

```
   Do not give the adviser credit for content stated only by the
      assistant.

Use only the supplied text. Do not use participant outcomes, outside
   information,
or assumptions about the adviser's intentions. The score represents the
    advice's
text-based potential effectiveness, not its proven causal effect.


SCORING SCALE

Score each of the five dimensions from 0 to 4:

0 = Absent or seriously deficient
1 = Weak
2 = Adequate but limited
3 = Strong
4 = Excellent

If the evidence falls between two scores, use the lower score unless
   the higher
anchor is clearly satisfied. Do not reward length by itself. Repetition
    is not
additional substance.


DIMENSION 1: RELEVANCE AND CONTEXTUAL FIT

Question: Does the advice address the recipient's actual goal and fit
   the context?

0 = Irrelevant, substantially misunderstands the problem, or works
   against the
    stated goal.
1 = Mentions the general topic but mostly gives generic commands or
   encouragement.
2 = Addresses the main goal but overlooks an important feature or
   constraint.
3 = Directly addresses the goal and important contextual constraints.
4 = Precisely targets the central problem, prioritizes what matters,
   and remains
    aligned with all important constraints or follow-up requests in the
        text.

Apply the same principle in both contexts:

- For quantitative-task advice, judge fit to efficient, accurate
```

```
   performance on
  the timed sequence task.
- For Casey advice, judge fit to the workplace problem, the recipient's
    careful
  work style, expressed concerns, and requests for help.


DIMENSION 2: ACTIONABILITY AND SPECIFICITY

Question: Does the advice tell the recipient what to do in concrete,
   observable
terms?

0 = Gives no usable action.
1 = Gives only vague imperatives, such as "focus," "try harder," "speak
    up," or
    "take the feedback seriously."
2 = Names at least one concrete action but leaves important execution
   details
    unspecified.
3 = Gives clear steps, examples, scripts, decision rules, or an ordered
    strategy
    that the recipient could readily try.
4 = Gives a precise and immediately executable procedure, including
   when or how
    to act and enough detail to minimize guesswork.

Concrete actions can differ by context. Examples include a search or
   look-ahead
procedure for the quantitative task and a structured update, meeting
   contribution,
or suggested wording in Casey. Judge the level of specificity, not the
   topic.


DIMENSION 3: RATIONALE AND STRATEGIC VALUE

Question: Does the advice explain why its recommendation should help
   achieve the
goal?

0 = Provides no rationale, or relies on reasoning that contradicts the
   advice.
1 = Merely asserts that the recommendation will help.
2 = Gives a plausible but broad or incomplete reason.
3 = Clearly connects the recommended action to the intended result
   through a
    sensible mechanism.
```

```
4 = Explains the mechanism clearly and identifies when, why, or under
    what
     conditions the strategy is useful, including a relevant tradeoff or
         check.

Do not use outside knowledge to decide whether a strategy is
    objectively correct.
Evaluate the reasoning and support present in the advice text itself.


DIMENSION 4: FEASIBILITY AND ADAPTABILITY

Question: Could the intended recipient realistically use the advice,
    and does it
accommodate obstacles or constraints revealed in the text?

0 = Impractical, unusable, or directly conflicts with a central
    constraint.
1 = Places substantial burden on the recipient or ignores obvious
    constraints.
2 = Generally feasible but generic, rigid, or only partly adapted.
3 = Realistic and appropriately adapted to the main constraints in the
    context.
4 = Realistic, low-burden, and flexible; offers a useful adjustment,
    fallback, or
     next step if the initial strategy does not work as expected.

For the quantitative task, relevant constraints include limited time
    and visual
search demands. For Casey, relevant constraints include workplace norms
    , the
recipient's careful style, discomfort with unfinished work, and the
    need to act on
feedback. Do not require emotional support in the quantitative context.


DIMENSION 5: CLARITY, COHERENCE, AND CALIBRATION

Question: Is the advice understandable, internally consistent,
    efficiently
expressed, and appropriately certain?

0 = Incoherent, centrally self-contradictory, or impossible to follow.
1 = Major ambiguity, disorganization, contradiction, or unsupported
    certainty
     substantially weakens the advice.
2 = Understandable overall but noticeably vague, repetitive,
    disorganized,
```

```
    inconsistent, or overstated.
3 = Clear, logically organized, internally consistent, and reasonably
   calibrated.
4 = Exceptionally clear and economical, with an easy-to-follow main
   strategy and
    careful distinction between a possibility, an observation, and a
       certainty.

Do not penalize ordinary speech fillers or transcription artifacts
   unless they
materially obstruct meaning. Treat categorical claims more cautiously
   when the
text gives no support or test for them.


TOTAL SCORE AND INTERPRETATION

Add the five dimension scores:

    Total score = Dimension 1 + Dimension 2 + Dimension 3
                  + Dimension 4 + Dimension 5

Maximum = 20 points.

Interpret the total as follows:

17-20 = Highly effective advice text
13-16 = Effective advice text
 9-12 = Mixed or moderately effective advice text
 5-8  = Weak advice text
 0-4  = Ineffective advice text

These categories describe text-based potential effectiveness only.


REQUIRED OUTPUT

Return one JSON object in the following format:

{
  "episode_id": "<provided identifier>",
  "context": "quantitative_task" or "casey_workplace",
  "scores": {
    "relevance_and_contextual_fit": {
      "score": 0,
      "evidence": "brief quotation or close paraphrase",
      "reason": "one-sentence justification"
    },
```

```
    "actionability_and_specificity": {
      "score": 0,
      "evidence": "brief quotation or close paraphrase",
      "reason": "one-sentence justification"
    },
    "rationale_and_strategic_value": {
      "score": 0,
      "evidence": "brief quotation or close paraphrase",
      "reason": "one-sentence justification"
    },
    "feasibility_and_adaptability": {
      "score": 0,
      "evidence": "brief quotation or close paraphrase",
      "reason": "one-sentence justification"
    },
    "clarity_coherence_and_calibration": {
      "score": 0,
      "evidence": "brief quotation or close paraphrase",
      "reason": "one-sentence justification"
    }
  },
  "total_score": 0,
  "effectiveness_category": "<category from the thresholds above>",
  "summary": "one sentence identifying the strongest feature and main
     weakness"
}

The total_score must equal the sum of the five integer dimension scores
   . Return
valid JSON only. Do not add dimensions, change the scale, or include
   commentary
outside the JSON object.
```

B.4.3 Prompt C

This evaluation rubric was not included in our preregistration but is a robustness check to the "common" evaluation rubric from Section B.4.2, this rubric is based on common evaluation criteria used in the communications literature known as Advice Response Theory (ART) (see, Feng & MacGeorge, 2010). The structure of the prompt is the same as that of Section B.4.2, only the content is different.

We scored the advice text on the following four dimensions: 'Response efficacy', 'Feasibility', 'Limitations management', and 'Confirmation and facework'. See the exact prompt at the end of this section. We score four dimensions, each on a 0–4 integer scale. 'Response efficacy' asks whether the advice communicates a plausible recommendation. 'Feasibility' asks whether the advice is understandable and implementable. 'Limitations management' asks whether the advice recognizes relevant obstacles. 'Confirmation and facework' asks whether the advice is

communicated taking into account the recipient's perspective while addressing the problem.

Adjudication was used two times out of 314 total evaluations.[A5] In the sequence task advice inter-rater agreement is high. Exact agreement between the two raters is 84% on response, 78% on feasibility, 72% on limitations, and 84% on confirmation, with quadratic-weighted kappas of 0.88 on response, 0.84 on feasibility, 0.67 on limitations, and 0.64 on confirmation. The two-rater mean has a Spearman–Brown reliability of 0.94. In the workplace advice inter-rater agreement is high. Exact agreement between the two raters is 89% on response, 87% on feasibility, 77% on limitations, and 85% on confirmation, with quadratic-weighted kappas of 0.85 on response, 0.88 on feasibility, 0.87 on limitations, and 0.93 on confirmation. The two-rater mean has a Spearman–Brown reliability of 0.98.

Table B2 contains the correlations between the rubric C mean score of advisor advice, advisor performance, and individual advisor characteristics. We can see positive but small correlations without statistical significance between both types of advice evaluated using rubric C and the advisor performance, yet a relatively much larger positive and statistically significant relationship between the advice of advisors evaluated using rubric C. This suggest that evaluation of advice without a ground truth element may miss key aspects of what advice leads to improved performance. We also find positive and statistically significant relationships between the sequence task advice evaluated with rubric C and an advisor's emotional perceptiveness and fluid intelligence.

```
    LLM PROMPT: COMMON ADVICE RESPONSE THEORY RUBRIC

You are evaluating the text-based quality and likely effectiveness of
   advice in
two different contexts:

1. Quantitative-task advice: advice about completing a specific timed
   number-
   sequence task. Each row in quant_advice_transcript.csv contains one
      advice
   text in advice_transcript.

2. Casey workplace advice: advice about responding to workplace
   feedback,
   becoming more proactive and visible, and doing so without abandoning
       a careful
   and thorough work style. In casey_transcripts.csv, "user" is the
      advice giver;
   "assistant_male" and "assistant_female" are the advice recipient.
      Evaluate all
   user turns for one prolific_pid together as one advice episode. Use
      assistant
   turns only to understand the recipient's situation, concerns, and
      requests.
   Do not give the adviser credit for content stated only by the
```

[A5] Each advisor has two evaluations here, one for each advice setting.

Table B2: Correlates of advisor performance with rubric C

	Advisor performance ($\hat{\mu}_i$) (1)	**Sequence advice rubric C** (2)
Advisor attributes[a]		
Emotion perception (PAGE)	–	0.24
		(< 0.01)
Fluid intelligence (Ravens)	–	0.24
		(< 0.01)
Personality mindset	–	0.03
		(0.73)
Intelligence mindset	–	-0.05
		(0.55)
Advice evaluation rubrics[b]		
Workplace advice rubric C	0.04	0.25
	(0.62)	(< 0.01)
Sequence advice rubric C	0.12	–
	(0.12)	

Notes: We report pairwise correlations of variables for all 157 advisors. In column (1) the dependent variable is the mean causal contribution of advisor i on advisee performance (see equation (2)) using our preregistered specification of model (1) i.e. advisor average causal contribution controlling for individual differences in baseline sequence task skill. In column (2) the dependent variable is the subjective evaluation of sequence advice based on common rubric C score derived from Feng and MacGeorge (2010) (for details on the rubrics see Section B.4 of the Appendix). [a]Advisor attributes see Section 1.4.1 for details. [b]Advice evaluation rubrics: See Section B.4 of the Appendix for details on rubric C. Workplace advice corresponds to the advice given in the workplace scenario (see Section 1.3) and Sequence advice corresponds to the sequence task advice (see Section 1.2). All variables have been standardized to have mean zero and standard deviation one. To express uncertainty, p-values are in parentheses. As stated in our preregistration our aim is to explore correlations and so we do not employ a multiple-hypothesis-test correction (Parker & Weir, 2022).

```
      assistant.

This rubric adapts four message-level constructs from Advice Response
   Theory for
external evaluation of advice text:

1. Response efficacy
2. Feasibility
3. Limitations management
4. Confirmation and facework

Advice Response Theory commonly measures recipients' perceptions of
   these
features. In this task, operationalize them as observable properties of
    the
supplied advice text. Do not infer how the recipient actually felt,
   whether the
recipient followed the advice, or whether the advice ultimately
   succeeded.

Use only the supplied transcript and task materials. Do not use
   participant
outcomes, outside information, or assumptions about the adviser's
   intentions.
Do not infer that an adviser is knowledgeable, trustworthy, likable, or
    similar
to the recipient unless evaluating a statement in the advice itself
   requires
doing so.

The scores represent the advice's text-based potential effectiveness,
   not its
proven causal effect.

SCORING SCALE

Score each of the four dimensions from 0 to 4:

0 = Absent or seriously deficient
1 = Weak
2 = Adequate but limited
3 = Strong
4 = Excellent

If the evidence falls between two scores, use the lower score unless
   the higher
anchor is clearly satisfied. Do not reward length by itself. Repetition
    is not
```

```
additional substance.

Evaluate each dimension independently. The same feature may be relevant
    to more
than one dimension only when it provides distinct evidence for each
   construct.

Relevance, specificity, clarity, coherence, and calibration are not
   separate
dimensions in this rubric. Consider them only insofar as they affect
   response
efficacy, feasibility, limitations management, or confirmation and
   facework.

DIMENSION 1: RESPONSE EFFICACY

Question: Does the advice communicate a recommendation that could
   plausibly
produce the recipient's desired result, and does it explain how or why
   the
recommended action should help?

0 = Provides no discernible recommendation, recommends something
   unrelated to
the stated goal, or presents reasoning that directly contradicts the
recommendation.

1 = Suggests only a broad action or desired outcome and gives little or
    no
indication of how the recommendation would help.

2 = Gives a recommendation that is plausibly connected to the goal, but
    the
connection is asserted, generic, or incompletely explained.

3 = Clearly connects a context-specific recommendation to the intended
   result
through a sensible mechanism or rationale contained in the text.

4 = Provides a clear and well-integrated account of how the key
   recommendation
should produce the intended result, identifies the most strategically
valuable action, and explains relevant conditions or priorities when
   needed.

A recommendation cannot receive a score above 2 merely because it
   sounds
sensible. Scores of 3 or 4 require the advice text to connect the
```

```
   recommended
action to the intended result.

Apply the same principle in both contexts:

* For quantitative-task advice, judge whether the adviser identifies a
   strategy
  for improving speed, accuracy, visual search, or overall performance
     and
  explains how the strategy should produce that improvement.

* For Casey advice, judge whether the adviser identifies behaviors that
    could
  improve proactivity, visibility, or responsiveness to feedback and
     explains
  how those behaviors could help while preserving careful and thorough
     work.

Do not use observed performance or outside knowledge to determine
   whether the
recommendation actually worked. Evaluate the proposed action-outcome
   connection
present in the supplied text.

DIMENSION 2: FEASIBILITY

Question: Could the intended recipient realistically understand and
   implement
the recommended action in the stated context?

0 = Provides no implementable action, proposes an action that is
   effectively
impossible in the stated context, or directly conflicts with a central
constraint.

1 = Gives a very vague instruction, places an unrealistic burden on the
recipient, or omits so much execution information that the recipient
   would
have substantial difficulty acting on it.

2 = Gives at least one generally usable action, but important execution
    details
are missing, or the recommendation is generic, rigid, or only partly
   adapted
to the context.

3 = Gives concrete and realistically usable guidance that fits the main
constraints and provides enough detail for the recipient to act.
```

```
4 = Gives precise and immediately executable guidance that is well
    adapted to
the context, including useful information about how, when, where, or in
     what
order to act when such information is needed.

Apply the same principle in both contexts:

* For quantitative-task advice, relevant evidence includes a concrete
    search,
  scanning, clicking, look-ahead, or prioritization procedure that can
      be used
  under limited time and visual-search demands.

* For Casey advice, relevant evidence includes realistic next steps,
    suggested
  wording, meeting behaviors, communication routines, timing,
      sequencing, or a
  plan for balancing greater visibility with careful work.

A concise recommendation can receive a high score when it contains
    everything
needed for implementation. Do not require unnecessary detail or
    multiple
recommendations.

DIMENSION 3: LIMITATIONS MANAGEMENT

Question: Is the recommendation free from important unaddressed
    drawbacks, and
does the advice recognize or manage relevant obstacles, risks,
    tradeoffs, or
uncertainties?

0 = The recommendation creates a serious problem, directly aggravates a
     central
constraint, or contains a major unaddressed contradiction that makes
    the
advice unusable.

1 = Ignores an obvious and important obstacle, cost, risk, or tradeoff
    revealed
in the supplied text.

2 = Is generally usable but leaves a material limitation underaddressed
    , treats
the recommendation as more universally applicable than the text
```

```
   supports,
or provides little guidance for handling a foreseeable obstacle.

3 = Avoids serious drawbacks and appropriately addresses the main
   limitation,
constraint, or tradeoff relevant to the recommendation.

4 = Handles the important limitations and tradeoffs in a well-
   calibrated way and,
when warranted, provides a useful adjustment, check, fallback,
   contingency,
or alternative.

Apply the same principle in both contexts:

* For quantitative-task advice, possible limitations include loss of
   time,
  excessive visual scanning, distraction, greater error risk, memory
     demands, or
  a strategy that may not apply uniformly across trials.

* For Casey advice, possible limitations include workplace norms, the
   risk of
  appearing performative or overcommitting, discomfort with sharing
     unfinished
  work, limited authority, competing workload demands, and tension
     between speed,
  visibility, and quality.

Do not reward boilerplate cautions. A limitation must be relevant to
   the advice
or context. If the supplied text reveals no material limitation
   requiring
discussion, judge whether the recommendation avoids creating an
   important
unaddressed drawback rather than requiring the adviser to invent one.

DIMENSION 4: CONFIRMATION AND FACEWORK

Question: Does the advice communicate respect for the recipient's
   competence,
perspective, identity, and autonomy while addressing the problem?

0 = Insults, humiliates, blames, coerces, or explicitly dismisses the
   recipient's
competence, perspective, or central concern.

1 = Uses noticeably patronizing, judgmental, dismissive, or
```

```
   unnecessarily
controlling language, or substantially disregards a concern expressed
   by the
recipient.

2 = Is basically civil but impersonal, somewhat blunt or overdirective,
    or offers
only generic positivity without meaningfully supporting the recipient's
competence, perspective, or choice.

3 = Is respectful and nonjudgmental, preserves reasonable autonomy, and
    responds
appropriately to the recipient's perspective when that perspective is
available.

4 = Skillfully acknowledges or validates the recipient's perspective,
   supports
competence and agency, and handles potential threats to identity or
   autonomy
without weakening the clarity of the recommendation.

Apply the same principle in both contexts:

* For quantitative-task advice, do not require emotional support,
   dialogue, or
  questions. Judge whether the advice is respectful, confidence-
     preserving, and
  non-condescending. A clear and respectful one-way message can score
     3. A score
  of 4 requires meaningful evidence that the adviser actively supports
     the
  recipient's competence or agency.

* For Casey advice, judge whether the adviser acknowledges the
   recipient's
  concerns, treats carefulness as a potentially valuable quality rather
      than a
  defect, respects the recipient's autonomy, and responds
     constructively to
  follow-up questions or reservations.

Questions, empathy, praise, or encouragement do not automatically earn
   a high
score. They count only when they meaningfully confirm the recipient's
   perspective
or protect competence, identity, or autonomy.

CROSS-CONTEXT EVALUATION RULES
```

```
Apply the same four constructs and scoring thresholds in both contexts.

* Do not penalize quantitative-task advice because the adviser had no
    opportunity
  for an interactive exchange.

* Do not automatically reward Casey advice for being interactive,
    longer,
  warmer, or more conversational.

* In Casey episodes, questions asked by the adviser count only when
    they help the
  adviser understand the problem or tailor subsequent advice.

* In Casey episodes, evaluate all adviser turns together. Advice may be
     developed,
  clarified, or revised across multiple turns.

* Use recipient turns only as contextual evidence. Do not attribute a
    recipient's
  ideas, explanations, or proposed actions to the adviser.

* If the adviser initially gives weak advice but later materially
    improves,
  clarifies, or corrects it, evaluate the complete episode. However,
      unresolved
  contradictions or harmful earlier guidance may still reduce the
      relevant
  score.

* Do not treat recipient agreement, gratitude, resistance, or stated
    intention to
  act as proof that the advice is effective or ineffective.

* Do not award points simply because the recommendation matches common
    sense.
  Scores must be supported by observable evidence in the supplied text.

TOTAL SCORE AND INTERPRETATION

Add the four dimension scores:

'''
Total score = Response efficacy
              + Feasibility
              + Limitations management
              + Confirmation and facework
```

```
'''

Maximum = 16 points.

Interpret the total as follows:

14-16 = Highly effective advice text
11-13 = Effective advice text
8-10 = Mixed or moderately effective advice text
4-7  = Weak advice text
0-3  = Ineffective advice text

These categories describe text-based potential effectiveness only.

REQUIRED OUTPUT

Return one JSON object in the following format:

{
"episode_id": "<provided identifier>",
"context": "quantitative_task" or "casey_workplace",
"scores": {
"response_efficacy": {
"score": 0,
"evidence": "brief quotation or close paraphrase",
"reason": "one-sentence justification"
},
"feasibility": {
"score": 0,
"evidence": "brief quotation or close paraphrase",
"reason": "one-sentence justification"
},
"limitations_management": {
"score": 0,
"evidence": "brief quotation or close paraphrase",
"reason": "one-sentence justification"
},
"confirmation_and_facework": {
"score": 0,
"evidence": "brief quotation or close paraphrase",
"reason": "one-sentence justification"
}
},
"total_score": 0,
"effectiveness_category": "<category from the thresholds above>",
"summary": "one sentence identifying the strongest feature and main
   weakness"
}
```

```
The total_score must equal the sum of the four integer dimension scores
   . Return
valid JSON only. Do not add dimensions, change the scale, or include
   commentary
outside the JSON object.
```

B.5 Heterogeneity Analysis

This section provides descriptive statistics and tests whether there are differences between genders within the roles of advisor or advisee. Table B3 contains the results of advisors and Table B4 contains the results for advisees. We can see that broadly speaking that within advisors the demographic characteristics are similar. In terms of experiment tasks, there is some evidence that females score marginally lower on the fluid intelligence test ($p = 0.08$, Holm $p = 1.0$). For advisees we find more statistically significant differences for relatively small differences, although that is to be expected with such large sample sizes, for example females tend to perform slightly worse at the sequence task than their male counterparts (Task 1:$p = 0.04$, Holm $p = 0.55$; task 2 $p < 0.01$, Holm $p < 0.01$; task 3 $p = 0.03$, Holm $p = 0.51$).

Table B3: Advisor Characteristics by Advisor Gender

Characteristic	Male ($N = 79$)	Female ($N = 76$)	Raw p-value	Global Holm-adjusted p-value
Age			0.036	0.538
18–24 years old (%)	10.1	7.9		
25–34 years old (%)	45.6	26.3		
35–44 years old (%)	16.5	35.5		
45–54 years old (%)	19.0	17.1		
55–64 years old (%)	6.3	11.8		
65–74 years old (%)	2.5	1.3		
Education			0.583	1.000
High school degree/GED (%)	7.6	11.8		
Some college (%)	27.8	30.3		
College degree or higher (%)	64.6	57.9		
Employment status			0.290	1.000
Part-time (%)	15.2	17.1		
Full-time employed (%)	60.8	46.1		
Self-employed or small business owner (%)	13.9	21.1		
Not employed (%)	10.1	15.8		
Perceived accent			0.033	0.535
Foreign (non-USA) (%)	19.0	5.3		
Regional USA (%)	6.3	7.9		
Standard USA (%)	74.7	86.8		
Born in the USA (%)	92.4	96.1	0.495	1.000
Race and ethnicity			0.369	1.000
Asian (%)	8.9	9.2		
Black or African American (%)	15.2	18.4		
Hispanic or Latino (%)	8.9	1.3		
Middle Eastern or North African (%)	1.3	1.3		
Native Hawaiian or Pacific Islander (%)	1.3	0.0		
White (%)	55.7	63.2		
Mixed or multiple races (%)	8.9	6.6		
Sequence task 1 score	10.7 (3.2)	10.8 (3.3)	0.936	1.000
Sequence task 2 score	13.6 (4.0)	13.2 (4.5)	0.517	1.000
Sequence task 3 score	13.8 (5.4)	12.7 (4.6)	0.172	1.000
Audio length (sec)	42.1 (23.1)	39.2 (22.3)	0.435	1.000
Fluid intelligence (Ravens)	5.8 (2.8)	5.0 (2.8)	0.083	1.000
Emotion perception (PAGE)	11.1 (2.6)	11.1 (2.2)	0.957	1.000
Personality mindset (Likert)	3.4 (1.2)	3.3 (1.1)	0.925	1.000
Intelligence mindset (Likert)	3.4 (1.1)	3.2 (1.1)	0.333	1.000

Notes: Categorical variables are reported as percentages. Continuous variables are reported as means with standard deviations in parentheses. The Other gender category is excluded. Likert indicates a numerical rating scale from 1 to 5. Race and ethnicity categories are mutually exclusive. Raw p-values were obtained using two-sided Welch tests for continuous variables and Pearson chi-squared tests for categorical variables. Global Holm-adjusted p-values jointly adjust all 23 advisor and advisee comparisons.

Table B4: Advisee Characteristics by Advisee Gender

Characteristic	Male ($N = 4{,}955$)	Female ($N = 5{,}005$)	Raw p-value	Global Holm-adjusted p-value
Age			0.010	0.179
18–24 years old (%)	11.2	10.7		
25–34 years old (%)	30.6	31.9		
35–44 years old (%)	26.7	23.8		
45–54 years old (%)	15.9	17.2		
55–64 years old (%)	10.4	11.5		
65–74 years old (%)	4.2	4.3		
75 years or older (%)	0.9	0.6		
Education			0.001	0.023
Less than high school (%)	1.2	0.9		
High school (%)	15.5	13.1		
Some college (%)	19.3	18.8		
College or higher (%)	64.0	67.2		
Employment status			< 0.001	< 0.001
Part-time (%)	12.3	19.0		
Full-time employed (%)	58.0	47.0		
Self-employed or small business owner (%)	11.4	13.6		
Not employed (%)	18.2	20.4		
Born in the USA (%)	36.8	37.0	0.859	1.000
Country of residence			< 0.001	< 0.001
United States (%)	40.7	40.3		
United Kingdom (%)	39.9	40.4		
South Africa (%)	3.0	5.6		
Australia (%)	2.6	2.6		
New Zealand (%)	0.7	0.4		
Ireland (%)	1.6	1.6		
Canada (%)	11.5	9.0		
Race and ethnicity			0.009	0.179
American Indian or Alaska Native (%)	0.2	0.3		
Asian (%)	12.8	10.8		
Black or African American (%)	15.8	16.9		
Hispanic or Latino (%)	2.9	2.2		
Middle Eastern or North African (%)	0.9	1.0		
Native Hawaiian or Pacific Islander (%)	0.1	0.1		
White (%)	61.1	61.9		
Other ethnic identity (%)	1.9	2.1		
Mixed or multiple races (%)	4.2	4.7		
Sequence task 1 score	9.6 (4.1)	9.4 (4.1)	0.038	0.538
Sequence task 2 score	12.6 (4.7)	12.2 (4.7)	< 0.001	0.003
Sequence task 3 score	11.3 (3.9)	11.1 (3.9)	0.030	0.506

Notes: Categorical variables are reported as percentages. Continuous variables are reported as means with standard deviations in parentheses. The Other gender category is excluded. Race and ethnicity categories are mutually exclusive. Raw p-values were obtained using two-sided Welch tests for continuous variables and Pearson chi-squared tests for categorical variables. Global Holm-adjusted p-values jointly adjust all 23 advisor and advisee comparisons.