Quantitative Assessment of Students' Revision Processes

Lisa R Volpatti, MIT

Lisa R. Volpatti is a Ph.D. candidate in the Anderson and Langer Labs at MIT with research interests in the development of responsive materials for biomedical applications. Prior to joining MIT, Lisa received her Masters of Philosophy in the Department of Chemistry at the University of Cambridge, UK and her Bachelor of Science in Chemical Engineering from the University of Pittsburgh. Lisa co-founded the Graduate Women in Chemical Engineering organization at MIT and is a NSF Graduate Research Fellow, a Whitaker International Fellow, and an MIT Chemical Engineering Communication Lab Fellow.

Mr. Alex Jordan Hanson, University of Texas at Austin
Jennifer M. Schall
Dr. Jesse N Dunietz, Massachussetts Institute of Technology

Jesse Dunietz is an educational designer for the MIT Communication Lab, an artificial intelligence researcher, and a freelance science writer. He develops training materials for the engineering graduate students who join the Communication Lab to serve as communication coaches for their peers. He holds a bachelor's in computer science from MIT and a Ph.D. in computer science from Carnegie Mellon University.

Amanda X Chen, Massachusetts Institute of Technology, Biological Engineering
Rohan Chitnis, Massachusetts Institute of Technology
Dr. Eric J. Alm
Dr. Alison F Takemura, US Department of Energy Joint Genome Institute

Alison loves wading into a good science story. Her first was her MIT doctoral thesis project, unlocking the gastronomical genome of a Vibrio bacterium. For some of the Vibrio's meals, she collected seaweed from the rocky, Atlantic coastline at low tide. (Occasionally, its waves swept her off her feet.) During grad school, Alison was also a fellow in MIT's Biological Engineering Communication Lab. Helping students share their science with their instructors and peers, she began to crave the ability to tell the stories of other scientists, and the marvels they discover, to a broader audience. So after graduating in 2015 with a microbiology doctorate, she trekked to the Pacific coast to study science communication at the University of California, Santa Cruz. There, she learned how to interview people, write feature stories, create podcasts, shoot videos, and finally, drive. Her stories were about pesticide residues in children, HIV in South Africa, rainforests in Australia, calorie-burning brown fat, and what hides behind Jupiter's clouds. Alison graduated in 2016, and like a homing pigeon, migrated back to MIT. There as the EECS Communication Lab manager, she supported the learning and growth of early scientists—eager to share their own stories. She now works at the US Department of Energy Joint Genome Institute, sharing science stories with an international audience.

Dr. Diana M. Chien, MIT School of Engineering Communication Lab

Dr. Diana Chien leads the MIT School of Engineering Communication Lab, and holds a PhD in Microbiology from MIT. Since 2013, she has coached, taught, and designed educational resources at multiple levels of the organization, including previous roles as a peer Communication Fellow, as the Biological Engineering Communication Lab manager, and as a Communication Instructor for undergraduate engineering courses. She is the co-founder of the CommKit, the Communication Lab's free online collection of discipline-specific guides to technical and professional communication. She is dedicated to promoting peer-to-peer professional development experiences for scientists and engineers.

Evaluating Peer Coaching in an Engineering Communication Lab: A Quantitative Assessment of Students' Revision Processes

Abstract

Communication is a crucial skillset for engineers, yet graduates [1]–[3] and their employers [4]–[8] continue to report their lack of preparation for effective communication upon completion of their undergraduate or graduate programs. Thus, technical communication training merits deeper investigation and creative solutions. At the 2017 ASEE Meeting, we introduced the MIT School of Engineering Communication Lab, a discipline-specific technical communication service that is akin to a writing center, but embedded within engineering departments [9]. By using the expertise of graduate student and postdoctoral peer coaches within a given discipline, the Communication Lab provides a scalable, content-aware solution with the benefits of just-in-time, one-on-one [10], and peer [11] training. When we first introduced this model, we offered easy-to-record metrics for the Communication Lab's effectiveness (such as usage statistics and student and faculty opinion surveys), as are commonly used to assess writing centers [12], [13].

Here we present a formal quantitative study of the effectiveness of Communication Lab coaching. We designed a pre-post test study for two related tasks: personal statements for applications to graduate school and graduate fellowships. We designed an analytic rubric with seven categories (strategic alignment, audience awareness, context, evidence, organization/flow, language mechanics, and visual impact) and tested it to ensure inter-rater reliability. Over one semester, we collected and anonymized 119 personal statement drafts from 47 unique Communication Lab clients across four different engineering departments. Peer coaches rubric-scored the drafts, and we developed a statistical model based on maximum likelihood to identify significant score changes in individual rubric categories across trajectories of sequential drafts. In addition, post-session surveys of clients and their peer coaches provided insight into clients' qualitative experiences during coaching sessions.

Taken together, our quantitative and qualitative findings suggest that our peer coaches are most effective in supporting the skills of organization/flow, strategic alignment, and providing appropriate evidence; this aligns with our program's emphasis on supporting high-level communication skills. Our results also suggest that a major factor in coaching efficacy is coach-client discussion of major takeaways from a session: rubric category scores were more likely to improve across a drafting trajectory when a category had been identified as a takeaway. Hence, we show quantitative evidence that through collaborative conversations, technical peer coaches can guide clients to identify and effectively revise key areas for improvement.

Finally, since we have gathered a sizable dataset and developed analytical tools, we have laid the groundwork for future quantitative writing assessments by both our program and others. We argue that although inter-rater variability poses a challenge, statistical methods and skill-based assessments of authentic communication tasks can provide both insights into student writing/revision ability and direction for improvement of communication resources.

Introduction

One of the greatest gaps in engineering education is the development of communication skills: degree accreditation agencies and employers alike identify communication as one of the most crucial skills [14]–[18], yet most graduates feel unprepared for the demands of professional communication [3], [18]. To fill this gap, educational programs have often adopted curricular interventions such as technical communication courses or embedded communication tasks within technical courses [19]–[21]. However, writing centers -- co-curricular interventions that provide students with just-in-time support throughout their training -- have been both underused and much less studied [9].

We previously introduced the Communication Lab (Comm Lab), an adaptation of the writing center model specifically for STEM contexts, which originated in 2012 in a single department at the Massachusetts Institute of Technology (MIT) [9], [22]. By training STEM graduate students and postdocs as peer coaches, the model leverages the educational benefits of peers' first-hand experience with communication in the discipline [23]–[26], learning through authentic tasks [27]–[29], and just-in-time support. We described the Comm Lab's original implementation within several MIT engineering departments in [9]. Subsequently, we compared its adaptations across several different technical and liberal-arts institutions in [22]. Our first publication underlined the affordability and flexibility of a peer coaching model, in contrast to a one-time curricular intervention. Likewise, our second publication highlighted the adaptability of the Comm Lab model to different institutional constraints and needs (e.g., service to undergraduates only *versus* both undergraduate and graduate students). Indeed, adaptation to local conditions is a core tenet of the model, and its success is demonstrated by the Comm Lab's continued growth across both MIT departments and external institutions.

The Communication Lab's core pedagogical approach

The Comm Lab's coaching model emphasizes self-analysis and incorporation of feedback through revision. An appointment with a Comm Lab coach encourages the client to take an active role in analyzing their work and proposing solutions; the coach facilitates by asking open-ended questions and acting as a proxy for the client's eventual, technical audience. A typical appointment of 30-60 minutes proceeds as follows: 1. The client and coach discuss the intended audience for the communication task and the client's own strategic goals. 2. The coach suggests an activity that will help the client analyze their own work (such as distilling the three most important points they wish to convey), while the coach reviews the work. 3. The coach focuses first on reviewing high-level communication choices like argument and structure, but also assesses the client's success in executing these according to field-specific expectations: e.g., is the logical flow of an argument technically sound? 4. Following assessment, the coach and client discuss the communication issues identified, compare examples from the field (which may include the coach's own experiences), and model/practice strategies for revision. 5. The coach ensures that the client identifies priorities for revision on their own. In short, during a session, the coach models for the client a process for both high-level analysis and practical revision.

Research on writing centers confirms numerous benefits of such peer learning experiences, including increased writer satisfaction, improved writing and revision processes, and improved course outcomes [30]. Empirical research likewise highlights the advantage of peers with disciplinary knowledge who can address both rhetoric and content by, for example, challenging students' technical claims and evidence [23]. In other words, a "knowledgeable peer" [31] offers a combination of social-emotional, communication, and technical support.

Our aims in designing a quantitative and qualitative study of the Communication Lab

In this study, our primary research question was: is the Comm Lab succeeding in improving clients' work according to our own metrics of success? I.e., do sessions bring clients closer to our standards for a given communication task, which are informed by both rhetorical principles and real-world field standards? To do so, we designed a quantitative, rubric-based, pre-post evaluation of authentic writing products: drafts for graduate school and graduate fellowship applications, assessed by authentic evaluators -- a team of our own peer coaches. In order to build a broader picture of the client's analytical and reflective experience, we complemented the quantitative core of the study by collecting qualitative reflections about the content of the coaching session. Overall, we argue that our study design provides useful qualitative and quantitative information about the effectiveness of the Comm Lab, despite the many limitations inherent in writing assessments.

Writing studies experts agree that writing assessments are challenging: whether quantitative or qualitative, of writing centers in particular or the writing process more broadly, it is difficult to design direct, authentic assessments that concretely demonstrate student success or growth [12], [32]. Our past publications [9], [22] offered typical indirect measures used by writing centers, such as repeat visits, client self-assessment, and faculty testimonials. While useful for program justification, such indirect metrics are clearly limited in their ability to concretely evaluate student growth [12], [13], [33].

Direct assessments are complicated by three considerations: validity, reliability, and ethical limitations on truly scientific study design. Validity asks: does the assessment measure what it is supposed to measure? Reliability asks: can writing be consistently and quantitatively evaluated by different evaluators? Finally, ethics forbid writing centers from executing the classic "treatment/no treatment" experimental design: true negative controls would require denial of writing center access to students who want it. Due to these three constraints, "the typical evaluation of writing programs...usually fails to obtain statistically significant results" [34]. For this reason, since roughly the 1990s, research on writing assessment and especially writing center assessment has focused on qualitative studies, despite the advantages of quantitative pre-post test design [26].

Nonetheless, we designed our study to maximize validity and reliability within these constraints by addressing the most important concerns and recommendations about both:

First, most concerns about validity revolve around the authenticity of the work: does the written response reflect how a student would do "in real life"? As Calfee and Miller observe, "best practices in writing assessment begin with an authentic task, where purpose and audience are clear and meaningful, [and] where support and feedback are readily available" [35]. While authenticity is often emulated through writing prompts, our study employs the truly authentic tasks of graduate school and fellowship applications, such that the written responses *are* the students' "real-life" writing performance.

Second, to achieve reliability, we used a small number of graders, calibrated against one another using an analytic rubric. The use of rubrics has generated a great deal of controversy, including critiques that rubrics often do not improve reliability [28], [29], that holistic scoring provides more validity and captures emergent properties of a written work [30], and that the "fear of disagreement" leads assessors to high-reliability, low-validity metrics with the illusory certainty that numbers imply [31]. However, defenders of rubrics argue their usefulness for combating bias, elucidating the properties of quality writing, and achieving consistent grading

when used properly [32]. With these considerations in mind, to execute our study, we designed a reliable and objective rubric through the thoughtful articulation of metrics, as discussed in the Methods. This design process included rigorous debate within the Communication Lab, ensuring that the rubric is broadly accepted to validly represent the salient features of successful writing.

Through these efforts to maximize both validity and reliability (expanded upon below), we aimed to provide useful evidence regarding the effectiveness of the Communication Lab, as well as analytical tools that can be adapted for future quantitative assessments..

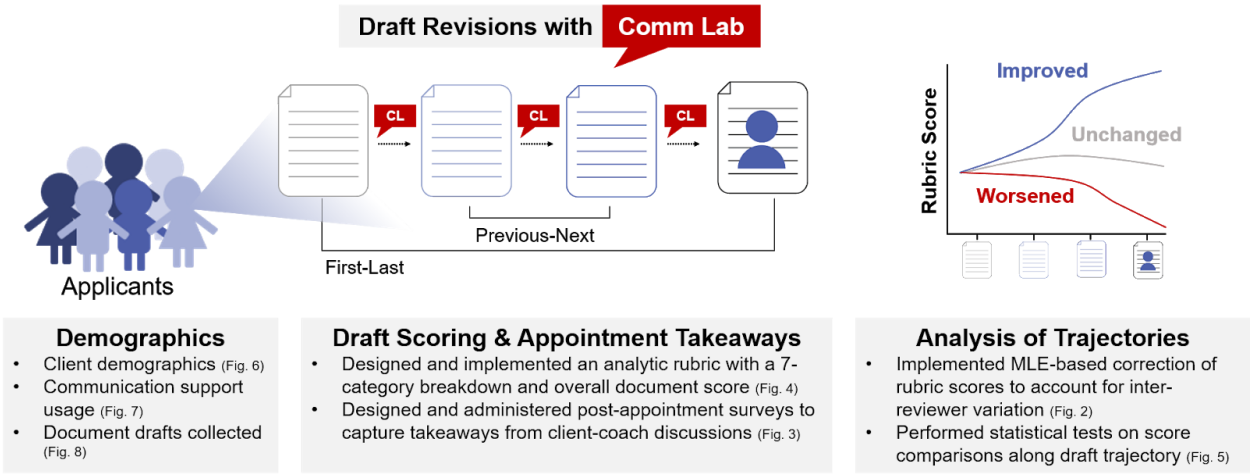


Figure 1: A qualitative and quantitative framework for assessment of writing center performance. Drafts of personal statements were collected from Comm Lab clients. Post-appointment surveys collected qualitative data including client demographics, communication support usage, draft characteristics, and takeaways from client-coach discussions. We developed an analytic rubric to enable quantitative assessment of document drafts. A team of peer coaches then applied the rubric to the drafts in a blinded fashion. Rubric scores were analyzed using a Maximum Likelihood Estimation-based (MLE) model to identify statistically significant changes along clients' trajectories of sequential drafts.

Methods

Overview of study design

We designed our study to explore the overall question of whether our communication coaching is making an impact on our clients' documents and drafting process, according to our own expectations for our peer coaches. The core questions to be answered were: do our peer coaches find improvement in client documents after they have been revised following a Comm Lab coaching session, and if so, how much and of what kind? Hence, we conducted a pre-post assessment of clients' drafts using an analytic rubric (i.e., a grid of scoring levels and categories for scoring and scoring levels) that we designed and tested for reliability. Figure 1 summarizes the study.

In brief, during the fall of 2018, we enlisted clients who were drafting applications to graduate schools or graduate fellowships, collected "trajectories" of sequential personal statement drafts, and rubric-scored the drafts. In addition, we collected survey data from both clients and their peer coaches. We aimed for these data to create a broader picture of the

clients' reflective and educational experiences during the coaching sessions, addressing questions such as whether the client and coach agreed upon the main takeaways and action items for improving a document. We chose to focus the study on personal statements (sometimes also called statements of purpose, statements of objectives, etc.) for several reasons: samples are abundant, clients are likely to return for multiple coaching sessions, and the document type more obviously showcases communication choices relative to technical/scientific propositions (as compared to, e.g., a grant proposal).

Our study proceeded in three phases: 1. collecting drafts and surveys, 2. scoring drafts according to our rubric, and 3. analyzing scores and survey responses (Figure 1). We elaborate on each below.

Methods 1: Draft and survey collection

The overall goal of data collection was to acquire drafts that were reviewed in coaching appointments for graduate school and fellowship personal statements, as well as clients' and coaches' surveyed reflections on the appointments. In addition, we collected final drafts (as submitted to the graduate school or fellowship), reflections, and participants' application success outcomes.

Consenting study participants were identified from among clients who signed up for personal statement coaching sessions in each of four MIT departmental Communication Labs (Biological Engineering, Chemical Engineering, Electrical Engineering and Computer Science, and Mechanical Engineering), during the fall of 2018. Gift cards were offered to incentivize survey completion. Surveys can be viewed in the Appendix. Intake client surveys included baseline questions such as primary language spoken and self-assessment of communication ability. Subsequent surveys for clients and coaches were required to be completed within one week of a coaching appointment and included questions about clients'/coaches' perceptions of the quality of the draft and communication topics discussed during the coaching session. Questions about communication topics were designed to parallel the categories in the rubric used to analyze drafts, so that we could assess whether coaching session takeaways resulted in improved rubric scores.

To collect final drafts and reflections, follow-up surveys were administered in January 2019, and a final survey to collect application outcomes was administered in May 2019.

We omitted from the study all "singleton" drafts, where the client submitted only one draft, and no others for comparison.

Methods 2: Draft scoring

Our rubric (see Appendix) was developed by a group of three peer coaches based on the National Science Foundation's Graduate Research Fellowship Program (NSF GRFP) guidelines for personal statements. The NSF GRFP was used as the basis because it is the most common graduate fellowship application at MIT. The initial NSF-focused rubric draft was broadened and adapted to cover personal statements for other fellowship programs and graduate school applications. The final rubric includes seven rubric categories based on the Communication Lab's criteria for effective professional communication, which reflect the classic rhetorical questions: 1. strategic alignment between the goals of the applicant and the fellowship/graduate school (why); 2. audience awareness (who); 3. appropriate context (what); 4. sufficient evidence (what); 5. logical organization/flow (how); 6. correct language mechanics (how); and 7. visual impact (how). (Visual impact refers to formatting choices that affect the reader's ability to skim the document, given that real application reviewers have limited reading time.)

The scoring for each category comprises four tiers: absent/incomplete (0), not competitive (1), somewhat competitive (2), and highly competitive (3). Thus, the maximum possible score for a full document was 21 points. Within each of the 7 categories, scores were assigned to two criteria, then averaged to obtain a category score. For example, the highly competitive tier for the audience category indicates that the document is 1. "easily comprehensible to the educated non-expert" and 2. "contains minimal instances of jargon." These criteria were chosen because typical review committees consist of

faculty members in the same field (e.g., chemical engineering) but not necessarily the same subfield (e.g., catalysis). The use of half points was discouraged within criteria, but could result from averaging scores across the two criteria within a category.

To improve inter-rater reliability, we created a glossary and user instructions for the rubric. For example, the glossary defines strategic alignment as "the ability of the document to address the review criteria and support the organization's mission." (If reviewers were not aware of review criteria provided by the specific graduate school or fellowship organization, they assumed criteria that broadly adhere to the NSF's guidelines of intellectual merit and broader impacts.) The user instructions ask the reviewers to evaluate the personal statement as MIT Communication Lab coaches rather than adopting the perspective of fellowship or graduate admissions committee members. With this mindset, the reviewers are to familiarize themselves with the rubric before an evaluation session. Reviewers are then instructed to skim each document for ~30 seconds and score the visual impact category of the rubric. Reviewers next spend 5-10 minutes reading the document in full before completing the remainder of the rubric. Finally, reviewers could also record a small amount of metadata about each draft: 1. flagging documents that were incomplete (e.g., ends mid-sentence or includes a phrase such as "add additional material here") and 2. optionally adding open-ended comments such as rationale for scores or potential biases regarding that particular document.

Prior to rubric scoring, all drafts were manually anonymized as follows: we removed all references to the person's name, names of individual faculty, names of projects, identifying information about publications, small companies/organizations worked for, and (when applicable/possible) names of current graduate programs. Where necessary to maintain coherence, such content was replaced with placeholders. No reviewers saw the un-anonymized drafts.

Scoring was performed by five peer coaches (including the three coaches who developed the rubric). To ensure inter-rater reliability, several rounds of testing were performed. Reviewers started by calibrating on singleton drafts that had been excluded from the study dataset. Once reviewers reached satisfactory agreement on the singletons, each proceeded to score 5-8 drafts from a shared test set, selected from the main dataset to represent a variety of stages of readiness and some full trajectories. In this latter test, reviewers almost always assigned full document scores that were within 2 points of each other. For all test documents except one, the standard deviation between scores was under 1.2 points; the one document with greater score variation received an outlier score that resulted in a standard deviation of 2.3. Based on these results, we calculated that a pair of reviewers would have about a 1-in-6 chance of differing on a given document by more than 2 points. We deemed this scoring consistency sufficient to proceed to the full study.

To prevent potential drift in scoring habits over the course of the scoring process, reviewers were instructed to re-read a set of two calibration drafts every ~15 documents. The calibration drafts were chosen from the testing set because they had high inter-rater agreement. The low calibration point draft received an average score of 12.5 with a standard deviation of 1.1 (n = 4), while the high calibration point draft received a score of 19 with a standard deviation of 0.8 (n = 3).

For scoring of the full set of documents, every draft was scored by at least two reviewers.[1] Drafts were assigned to reviewers automatically via a constrained optimization procedure as follows: 1. No reviewer could score a draft by a client whom they coached at any point during the semester (a hard constraint). 2. Each client's drafts should be distributed among the reviewers (i.e., a single reviewer should score a second or third draft from the same client as rarely as possible.) 3. For each draft, the number of reviewers from the same department as the author should be as close to 1 as possible (i.e., each draft should ideally be scored by one reviewer from the same department as the author and one reviewer from a different department.) Reviewers had access to information about which fellowships/graduate schools for which each document was intended, in order to adjust their reviewing expectations based on varying application prompts.

[1] Drafts on which reviewers iterated together while developing the rubric were scored by as many of the reviewers as possible, i.e., up to five. Beyond this set, we aimed for exactly two reviewers per draft. However, after scoring we realized that a few of the drafts had been erroneously submitted as multiple copies, which had been scored separately by multiple reviewers. This provided a further check on consistency: no reviewer disagreed with themselves on the total score by more than 1.5 points out of 21 total, and most were within 0.5 points.

For each document, we report rubric scores averaged across all reviewers.

Methods 3: Data analysis

To allow quantitative analysis of clients' and coaches' qualitative survey data about appointment experiences, Likert-scale responses were converted to numerical values (e.g., "not at all," "a little," a lot" were converted to 0, 1, 2). Surveys also included a free-response question in which clients and coaches summarized the appointment's "biggest takeaway" (i.e., most important skill discussed). These responses were analyzed in two ways. First, to assess client-coach agreement, two peer coaches compared the takeaways for each appointment and assigned a client-coach match score of 0 (completely different ideas), 1 (same general idea), or 2 (exact match). We report the averaged match scores. Second, the coaches categorized the takeaways according to the seven categories from the analytic rubric. In many cases, the takeaways fell under multiple rubric categories. To account for coaches' variations in categorization, we considered a takeaway as mentioning a given rubric category if at least one of the two coaches identified it as such.

To analyze clients' drafting progression based on our rubric scores, each client's trajectory of sequential fellowship drafts was curated to separate out trajectories for specific fellowships. E.g., a trajectory including both NSF GRFP and National Defense Science and Engineering Graduate (NDSEG) Fellowship drafts would be split into two trajectories: one for each of the two fellowships. Graduate school trajectories were not curated in this manner because graduate school application prompts tend to be more similar than fellowships'.

Although the rubric is numeric, we cannot assume that the scores are linear and can be averaged using a simple arithmetic mean. For example, scores of 2.0 and 3.0 from two independent reviewers might not be equivalent to scores of 2.5 and 2.5 from the same reviewers. Therefore, to identify statistically significant changes in rubric category score across trajectories, we developed a maximum likelihood estimation (MLE) scoring model that could account for arbitrary and non-linear effects. In short, the model statistically infers a consensus score among multiple reviewers; we call this the "true score."

In this model, each rubric category for each document was treated as independent, and modeled as having a score of 1, 2, or 3. Based on our reviewers' scores, the model assigned a probability that each rubric category's score was actually 1, 2, or 3. The highest-probability score was called the true score. (While we also tested allowing a possible true score of 0, which reviewers used to indicate missing content, the resulting inconsistencies led us to omit this possible score in our analyses.)

To train the MLE model, we used a set of eight documents scored by all five reviewers. We used the Expectation-Maximization (EM) algorithm [36] to calculate the probability that, from this training set, each rubric category with a true score of 1, 2, or 3 would receive a score of 0, 0.5, 1, 1.5, 2, 2.5, or 3 from a given reviewer. We compared the inferred true scores with reviewers' scores and found good agreement: reviewers were most likely to assign scores that agreed exactly with the true scores (Figure 2). Higher true scores (2 or 3) had a tighter distribution of reviewer scores. This could be explained by two causes: first, there is simply less scoring space possible around the maximum score of 3; and second, reviewers react more variably to weaker writing. Indeed, reviewers reported this trend anecdotally during the reviewing process. Overall, though, we took the strong agreement between true and reviewer scores as another indication that our reviewers had achieved consistent use of the rubric, despite the limitations of inter-rater variability.

The MLE model also allowed us to detect statistically significant changes within a client's trajectory of sequential drafts. Because we knew the probability of each score for each rubric category, we could compare that score across two drafts and compute the probability that a document had improved, worsened or stayed the same. We used a probability cutoff of 0.5 to call significant score changes: if a score neither improved nor worsened with $p \geq 0.5$, we considered it unchanged.

Figure 2: Reviewers are most likely to assign rubric category scores that agree with the scores statistically inferred using a Maximum Likelihood Estimate model. Reviewers are more consistent with the MLE-inferred score for higher scores. Heatmap values indicate the probability that a given inferred score was assigned a given score by a reviewer.

Results and Discussion

Overview

In this study, we performed both quantitative and qualitative analyses of the drafting process of MIT Communication Lab clients working on fellowship and graduate school applications. To do so, we collected personal statement drafts and qualitative surveys about the Comm Lab coaching experience and performed quantitative pre-post analysis of the personal statements using an analytic rubric. The qualitative survey data allowed us to characterize demographics and behavior of this set of Comm Lab users. With the quantitative data, we aimed to assess which communication skills were more or less addressed during these coaching appointments, assess connections with client outcomes, and translate findings into recommendations for improving the Comm Lab's efficacy. At a methodological level, we also aimed to inform future assessments by investigating the utility of different quantitative analytical methods, given the notorious difficulty of quantitative writing center assessments (as discussed in the Introduction).

In total, we collected data from 47 clients: 81 drafts reviewed during appointments, 38 final drafts, 108 post-appointment surveys (collecting demographic information and reflections about the coaching experience), and data about the outcomes of 86 applications. In addition, to capture the peer coaches' perspective, we collected 99 post-appointment surveys from 28 coaches. We note that our data are not necessarily representative of the general MIT population, since they include only students from four engineering departments who chose to 1. schedule Comm Lab appointments for fellowship/graduate school personal statements and 2. participate in the study [37]. In our discussions, we indicate potentially nonrepresentative traits of the study population through comparison to data about the general population where possible.

Below, we describe our results by first focusing on the content and impact of coaching sessions: topics discussed by coaches and clients (Results 1), the rubric scores assigned by our reviewers (Results 2), and our statistical analysis of score changes during the drafting process (Results 3). We end with more detailed characterization of clients (Results 4) and the collected documents (Results 5).

Results 1: Characteristics of Communication Lab appointments

To characterize the content and efficacy of the client-coach discussions that take place during Comm Lab appointments, our post-appointment surveys collected qualitative reflections from both clients and coaches. One major goal was to investigate the extent to which clients are able to distill the analytical skills and practical revision approaches discussed during the coaching session into a single message. A complementary goal was to investigate how effectively coaches collaborated with their clients in explaining or supporting these skills. Accordingly, we asked both clients and coaches to provide a free response where they distilled the "biggest takeaway (e.g., most important skill discussed)" from a coaching appointment (Figure 3).

First we investigated the question: are coaches effectively communicating the greatest priorities for revision to their clients? We analyzed the takeaways for client-coach agreement using two approaches: 1. hand-scoring them for agreement on a quantitative scale from 0 (completely different) to 2 (exact match) (Figure 3a), and 2. hand-scoring them for mentions of each of the seven skill categories from our rubric (Figure 3b). We found that overall client-coach agreement was high: the largest fraction of takeaways were scored as exact matches (49%), and overall, 86% of takeaways had some degree of agreement (Figure 3a). As an example of an exact match, one client reported that their takeaway was "Adding more emphasis to the thesis statement of my application - making it clear and explicit," while their coach reported, "Framing thesis of personal statement. Making statements clear and concrete." Analyzing the takeaways based on mention of rubric categories also found strong agreement: the aggregate distributions of communication skills discussed according to clients and coaches were extremely similar (Figure 3b). The largest discrepancy is in organization/flow, which clients included as a major takeaway more often than coaches.

This analysis of client and coach takeaways categorized by skills also demonstrated that the majority of Comm Lab appointments focus on high-level communication skills rather than mechanics, which is consistent with program goals. The most-discussed skills were organization/flow, strategic alignment, audience alignment, and use of evidence. The least-discussed were language mechanics, context, and visual impact (Figure 3b). Overall, these skill distributions appear consistent with the needs of persuasive writing. For example, the fact that takeaways involving organization/flow and strategic alignment outrank evidence suggests that Comm Lab clients are fairly successful in reporting adequate evidence of their qualifications for an application; however, they need to work on arranging that evidence into a coherent narrative that clearly expresses their alignment with the graduate school or fellowship. For the purposes of program improvement, these findings suggest that training curriculum for Comm Lab peer coaches might further incorporate strategies and exercises to help coaches work with clients on organization/flow in particular.

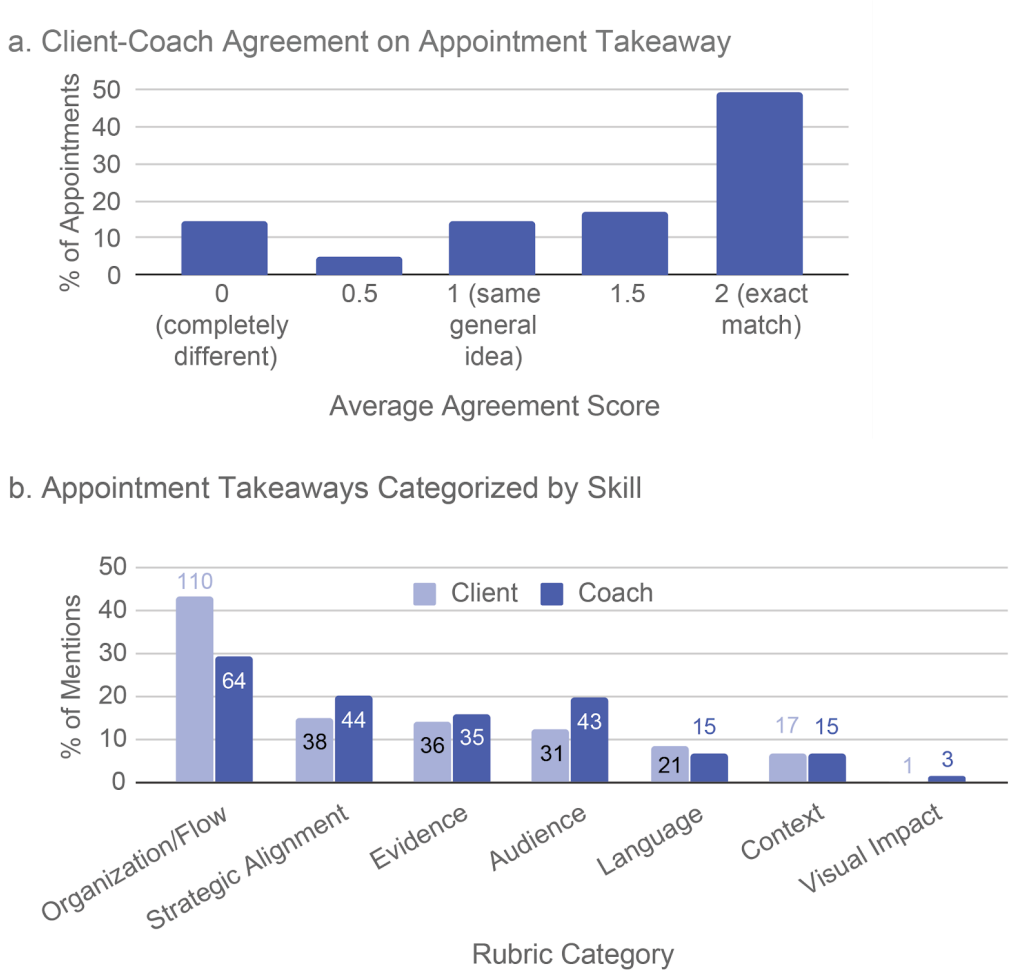


Figure 3: Clients and coaches usually agreed about the main takeaway from Comm Lab appointments, and takeaways were most often high-level communication skills rather than mechanics. **a)** Agreement in the content of client and coach survey responses about appointment takeaways: averaged scores for takeaway agreement from two reviewers (0 = completely different ideas, 1 = same general idea, 2 = exact match). **b)** Takeaways scored for mentions of the 7 skill categories from our rubric (union of scores from two reviewers). Bar labels are absolute counts.

Results 2: Overview of rubric scoring results

A trained team of peer coaches scored the 119 collected documents using the analytic rubric developed for this study (see Methods 2). The scores reported are averages of at least 2 reviewers, with a maximum score of 21 points total: 3 points per each of 7 categories (Figure 4). The distribution of scores is skewed toward higher scores, i.e., between “somewhat competitive” and “highly competitive” in our rubric (Figure 4a). Across all fellowship and graduate school documents, the mean total score is 15.6 points and the median is 15.8 points (Figure 4a). The relatively high scores appear consistent with two other features of the data set: the majority of participating clients rated themselves as “good” writers (discussed further in Results 4, Figure 6d), and the majority of documents collected were later-stage drafts (discussed further in Results 5, Figure 8c).

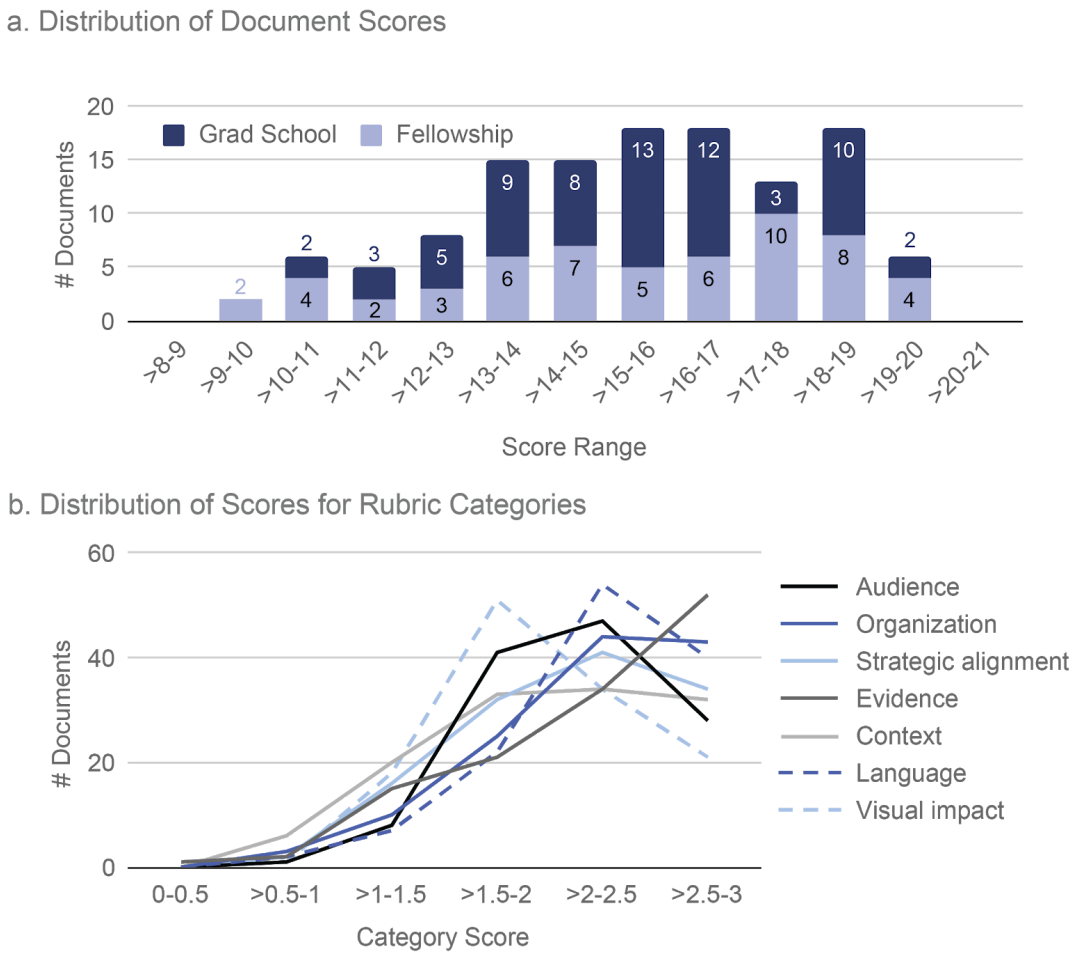


Figure 4: Scores for documents overall and individual rubric categories are skewed toward higher values. Distribution of rubric scores averaged from two reviewers for **a)** overall documents, separated by document type (overall mean = 15.6) , and **b)** rubric categories representing high-level skill categories (solid line) and low-level skill (dashed line) categories had an overall mean of 2.2.

The lowest scores present (9-11 points) were mainly for fellowships, particularly documents intended for the NDSEG and GEM fellowships. Three hypotheses may explain this, and can be investigated in the future: 1. Clients may have used their NSF personal statements (due late October) as a starting point for later fellowships (e.g., NDSEG, due early December), and these "starting points" may not yet have been substantially updated to align with the later fellowships' guidelines when submitted to the Comm Lab. 2. Study reviewers might have been more accustomed to evaluating NSF personal statements and struggled to adjust expectations to other fellowship types. For example, NDSEG statements might appear lacking in evidence compared to NSF ones because of their far shorter length, 500 words *versus* three pages. 3. The rubric might be biased because NSF expectations (e.g., inclusion of "broader impacts") were used as its initial basis. This low-scoring bias for non-NSF fellowships underlines anecdotal reports from Comm Lab peer coaches that they may feel less prepared to support applications to less common fellowships. To better prepare peer coaches for such appointment types, it may be valuable to provide them with access to additional examples and guidelines for such fellowships.

The predominance of high scores is also apparent when analyzing individual rubric categories (Figure 4b). The mean per-category score was 2.2 out of 3. Listing the categories from most left- to right-skewed distributions, their median scores were as follows: visual impact,

2.0; context, 2.2; strategic alignment, 2.3; audience, 2.4; organization/flow, 2.4; language mechanics, 2.5; and evidence, 2.5. The fact that evidence is the highest-scoring category reflects the culture of MIT, which promotes independent undergraduate research and leadership experiences. By contrast, the fact that visual impact is the lowest-scoring category reflects many clients' lack of awareness that they can use small visual formatting choices such as bolding and headers to improve reviewers' ability to skim. However, visual formatting was the least-frequently addressed appointment takeaway category (Figure 3b), likely because formatting choices can be discussed and revised quickly (and formatting is ultimately less significant than content for personal statements).

Interestingly, we previously noted that organization/flow was the most frequently addressed appointment takeaway category (Figure 3b), yet it was also one of the higher-scoring categories (Figure 4b). One possible explanation for this finding is that clients with self-evaluated good-to-excellent writing ability generally write in a logical progression, and additional modifications to the flow of their arguments may not result in substantial score increases. At the same time, since organizational decisions affect multiple areas of the document, discussions about organization tend to require more lengthy, in-depth conversations (compared to e.g. visual formatting) to explore different options.

Overall, since rubric scores for both full documents and individual categories were higher than anticipated, a future study focusing on a population with lower initial scores would be a useful comparison. To maximize the utility of this dataset, however, our next analysis focuses on a subset of rubric category scores that allows the most informative pre-post analysis.

Results 3: Maximum Likelihood Estimation analysis of rubric scores

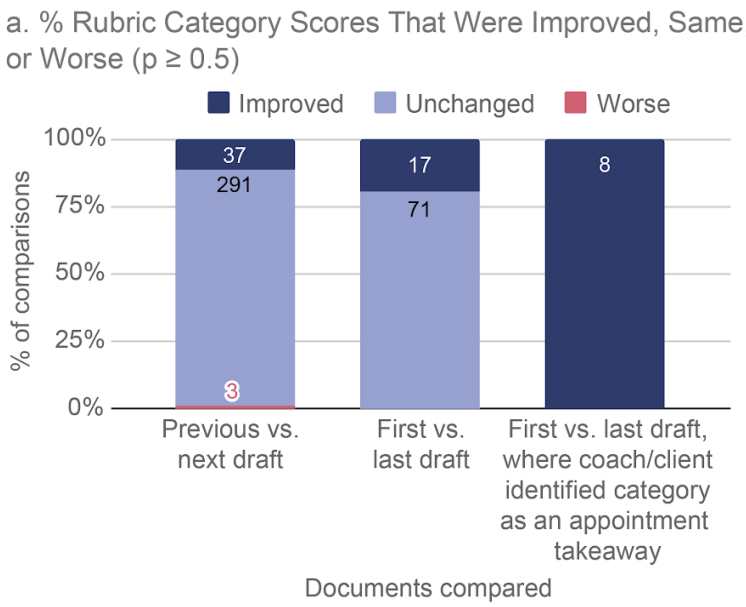
We aimed to analyze rubric scores to quantitatively assess documents' changes over the drafting process. However, because we could not control for all variation among reviewers, quantitative analysis of the dataset was challenging: scoring noise prevented us from detecting clear changes in the scores for clients' sequential drafts. This was especially the case because the drafts were skewed toward higher scores, and hence had little room for growth.

To address this noise, we developed a statistical approach to identifying a consensus score based on reviewers' scores. We focused this analysis on scores at the level of rubric categories, to provide insight into changes in individual skill areas. In short, we developed a model based on Maximum Likelihood Estimation (MLE), which used reviewers' rubric category scores to determine a consensus score (see Methods 3 for detail). After the model was trained on documents scored by all five reviewers, it was able to infer the most statistically probable "true score" for input scoring data, hence correcting for inter-reviewer variability. The model also identified statistically significant score increases and decreases within clients' drafting trajectories. Score changes with a probability ≥ 0.5 were considered significant; otherwise, scores were considered unchanged. We focused the analysis of significant score changes on an informative subset of data: we analyzed only comparisons where the initial category was statistically inferred to have a true score of 2 points. This allowed us to detect increased, decreased, or unchanged scores. (By contrast, an initial category score of 3 would allow us to detect only decreased scores, for example.)

Figure 5 summarizes findings from this analysis, using two types of category comparisons within drafting trajectories: comparing previous-next draft pairs (which allowed a greater number of comparisons), or comparing only first-last draft pairs. Inspection of these data revealed that the vast majority of rubric category scores were either unchanged or improved across a draft trajectory (Figure 5a). We observed that substantially more previous-next

comparisons were statistically inferred to have improved (11%) than worsened (<1%; worsened categories could be accounted for by intermediate stages in the drafting process). The percentage of improved scores increased as we focused on increasingly meaningful subsets of the data: first-last comparison, and comparisons where the coach and/or client had identified the rubric category in question as an appointment takeaway. We observed that 19% (n=17) and 100% (n=8) of these comparisons had significant improvement, respectively. This suggests that coach/client identification of an area of focus correlates with greater revision efficacy. The large majority of unchanged scores (88% of previous-next and 81% of first-last comparisons) indicates that our coaches have further potential to help their clients improve their work, even in this relatively high-performing population.

Finally, we investigated the frequency of improvement in each specific skill area (Figure 5b). Though interpretation of this analysis is limited by small sample size, the top three most frequently improved categories were the same for both the previous-next and first-last comparisons: language mechanics, evidence, and organization/flow. The category distribution was also similar to those previously identified through the analysis of clients' and coaches' appointment takeaways (Figure 3b), notably in the high ranking of evidence, organization/flow, and (for previous-next comparisons) strategic alignment. By contrast, language mechanics were highly represented among significant score increases, but little represented among appointment takeaways (Figure 3b). A potential explanation is that clients are able to significantly improve language mechanics through independent revision or using other support sources. Overall, the analysis of significantly improved skill areas again suggests that appointment takeaways and improvement are correlated, and that Comm Lab coaching is associated with improvement of higher-level communication skills.



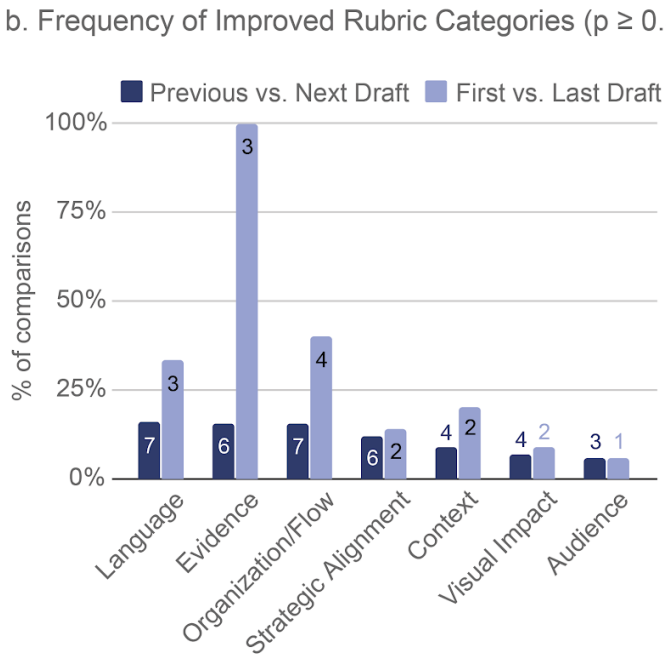


Figure 5: Within clients' drafting trajectories, nearly all rubric category scores were unchanged or improved, and were more likely to improve if coaches and/or clients had identified the category as an appointment takeaway. Rubric category score changes were statistically analyzed via a Maximum Likelihood Estimation model. Analysis focused only on score comparisons where the initial rubric category score was 2 out of 3 total points, and used a probability cutoff of 0.5 to call changed vs. unchanged scores. **a)** Distribution of statistically significantly improved, unchanged, or worsened rubric category scores for different comparisons within draft trajectories. **b)** Distribution of significantly improved rubric categories, as a percentage of previous-next or first-last comparisons. Bar labels are absolute counts.

Results 4: Client demographics and behavior

Figure 1 summarizes client demographics. The majority of clients were undergraduate (45%) and early graduate students (36% first year, 13% second year) (Figure 6a). Clients'

department affiliations were Biological Engineering, Chemical Engineering, Electrical Engineering and Computer Science, and Mechanical Engineering, with the largest fraction (39%) coming from Chemical Engineering (Figure 6b). With respect to English fluency, 77% of clients self-identified as native English speakers (Figure 6c). Finally, in a self-assessment of their writing ability, the majority of clients (70%) identified as “good” on a 4-point scale including “poor,” “fair,” “good,” and “excellent” (Figure 6d).

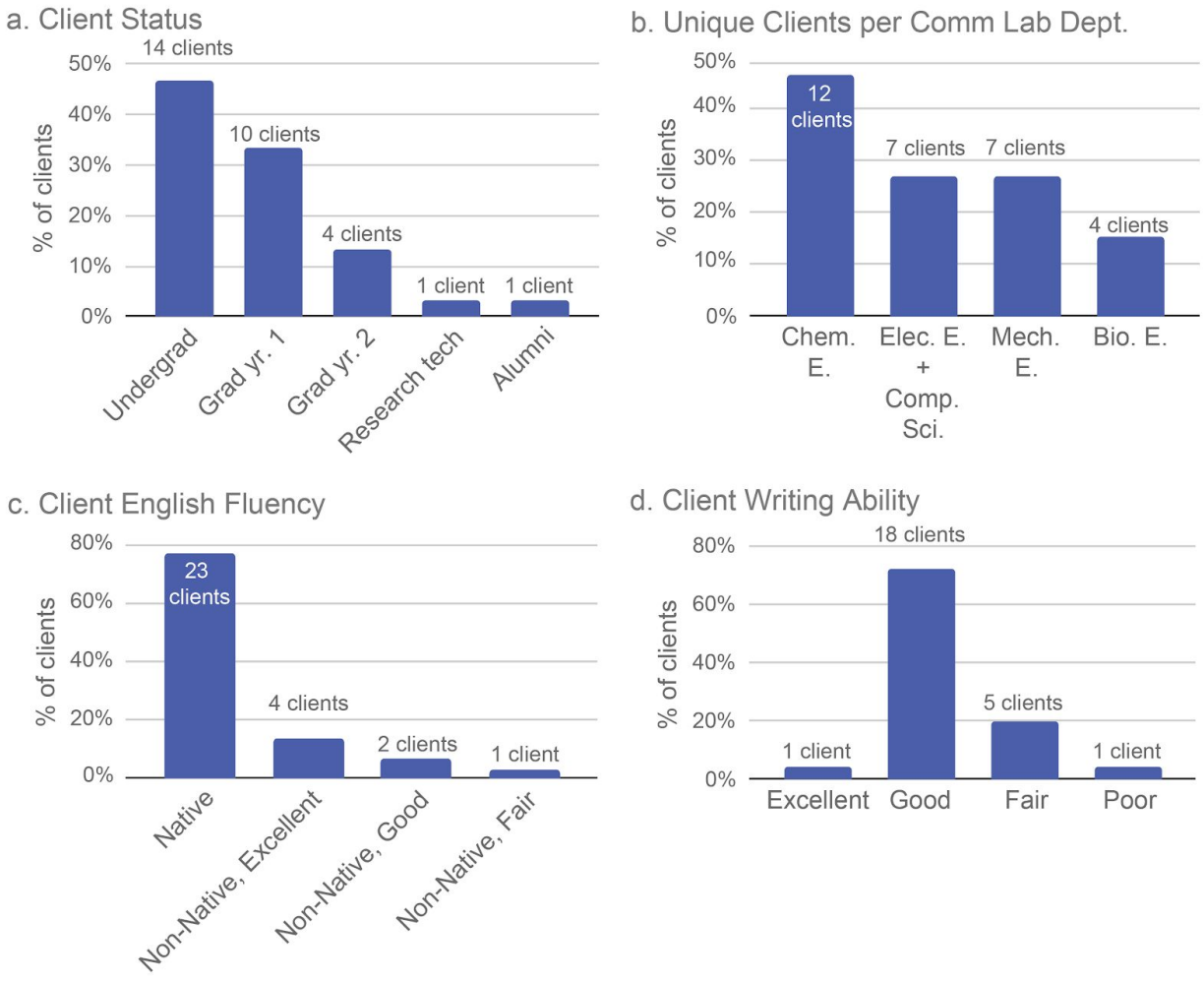


Figure 6: Overview of client demographics, including **a)** status at MIT, **b)** department affiliation, **c)** self-assessment of English fluency, and **d)** self-assessment of writing ability (missing data from 5 clients). Bar labels are absolute counts.

In Figure 7, we summarize clients’ habits in seeking writing support from services and personal/professional contacts. Clients most often made 1-2 Comm Lab appointments (8 clients each), followed by 4 appointments (7 clients, Figure 7a). However, several clients visited the Comm Lab upwards of 4 times, with one instance of 10 appointments (Figure 7a). We also asked clients to report the number of hours spent consulting other support sources: friends/family; undergraduates; graduate students/postdocs; faculty; or the MIT Writing and Communication Center, a centralized resource offering 1:1 appointments with writing experts rather than peer coaches (Figure 7b). The average client used a total of 12.1 hours of support, with the top 3 support sources being consulted almost equally: friends/family, Comm Lab peer coaches, and undergrads (3, 2.9, and 2.6 hours respectively). Overall, these data are consistent

with anecdotal student reports that students are most comfortable seeking support from personal contacts and peers, particularly ones with direct personal experience in the relevant communication genre. Hence, the Comm Lab coaches' experiential knowledge is a strong match for this preference: all have recently succeeded in graduate school applications in similar disciplines to their clients, and many have applied to and been awarded fellowships. To contextualize these findings about time spent seeking feedback, we note that the study population is more likely to seek feedback than the general population: previous Comm Lab surveys (unpublished) indicate that Comm Lab non-users often are working up to the last minute, dislike asking for help, and/or feel confident working independently.

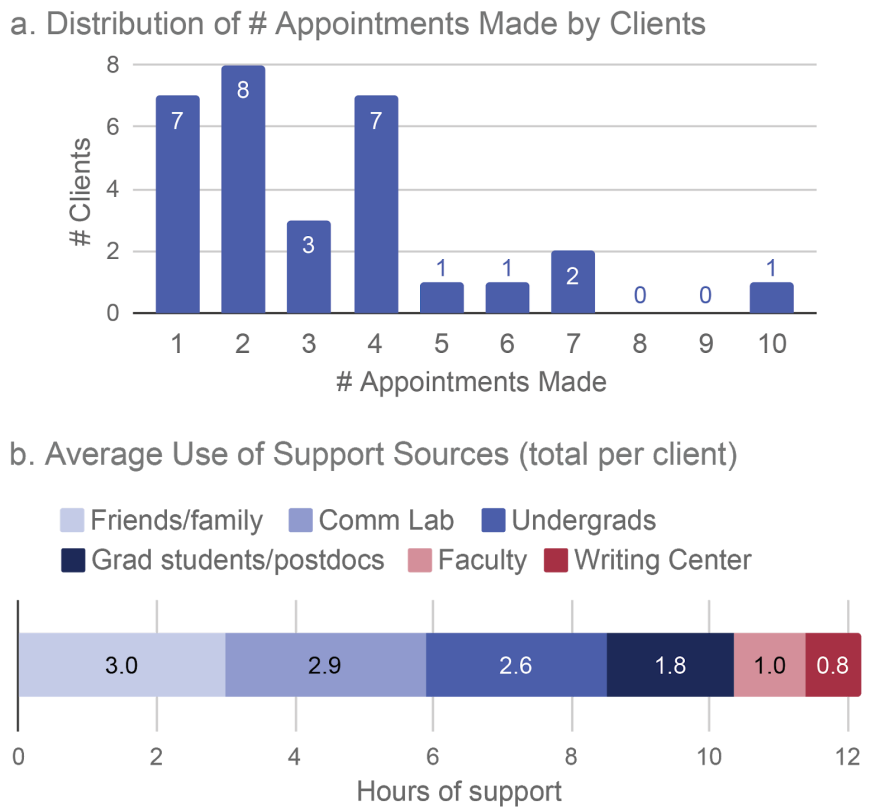


Figure 7: Overview of client use of communication support, including **a)** the distribution of number of Comm Lab appointments made by different clients, and **b)** average hourly use of the Comm Lab and other communication support resources, totaling 12.1 h. "Writing Ctr" indicates the MIT Writing and Communication Center, a centralized resource that offers 1:1 appointments with writing experts rather than peer coaches. Bar labels are absolute counts.

Results 5: Characteristics of collected personal statements

Figure 8 summarizes the characteristics of collected personal statements, as reported in surveys. Among the documents collected, fellowship and graduate school applications were represented almost equally: 45% and 55%, respectively (Figure 8a). Among fellowship personal statements, the majority were for the NSF GRFP (47%), NDSEG (28%), and Hertz Fellowship (14%) (Figure 8b). Clients also characterized each document's stage in the writing process. Almost all clients used the Comm Lab for documents that they considered to be rough drafts, polished drafts, or ready to submit to the application, with only a single document identified as being in the "ideas" (brainstorming) or outline stage (Figure 8c). This bias toward late-stage documents is consistent with generally observed and self-reported Comm Lab client behavior:

clients tend to be too self-conscious or last-minute to seek coaching during earlier stages. Consequently, it is frequently a strategic priority for the program to encourage earlier-stage Comm Lab appointments, in order to permit deeper discussions and planning of communication choices rather than later-stage edits.

Finally, clients reported their application outcomes (Figure 8d). We analyzed success rates from the perspectives of both success per application (was an individual application successful?) and success per client (did an individual client have at least one successful application?). The per-application success rates for fellowships and graduate school applications were 37% and 71%, respectively (Figure 8d). The fellowship success rate of 37% can be contextualized by comparisons to publicly available statistics: while the national average acceptance rate for the NSF GRFP is approximately 15%, an analysis of the 2019 award winners indicates that GRFPs are disproportionately awarded to certain institutions, including UC Berkeley, Stanford, and MIT [38]. Since most clients applied to multiple fellowships and graduate schools, we also present the data in terms of outcomes from a per-client perspective: 55% of clients applying to fellowships and 92% of clients applying to graduate programs had at least one successful application (Figure 8e). Although departments do not systematically collect data about students' application success rates for comparison, these data may provide a useful basis for future assessment of the Comm Lab, e.g., through comparison to a control population of non-users.

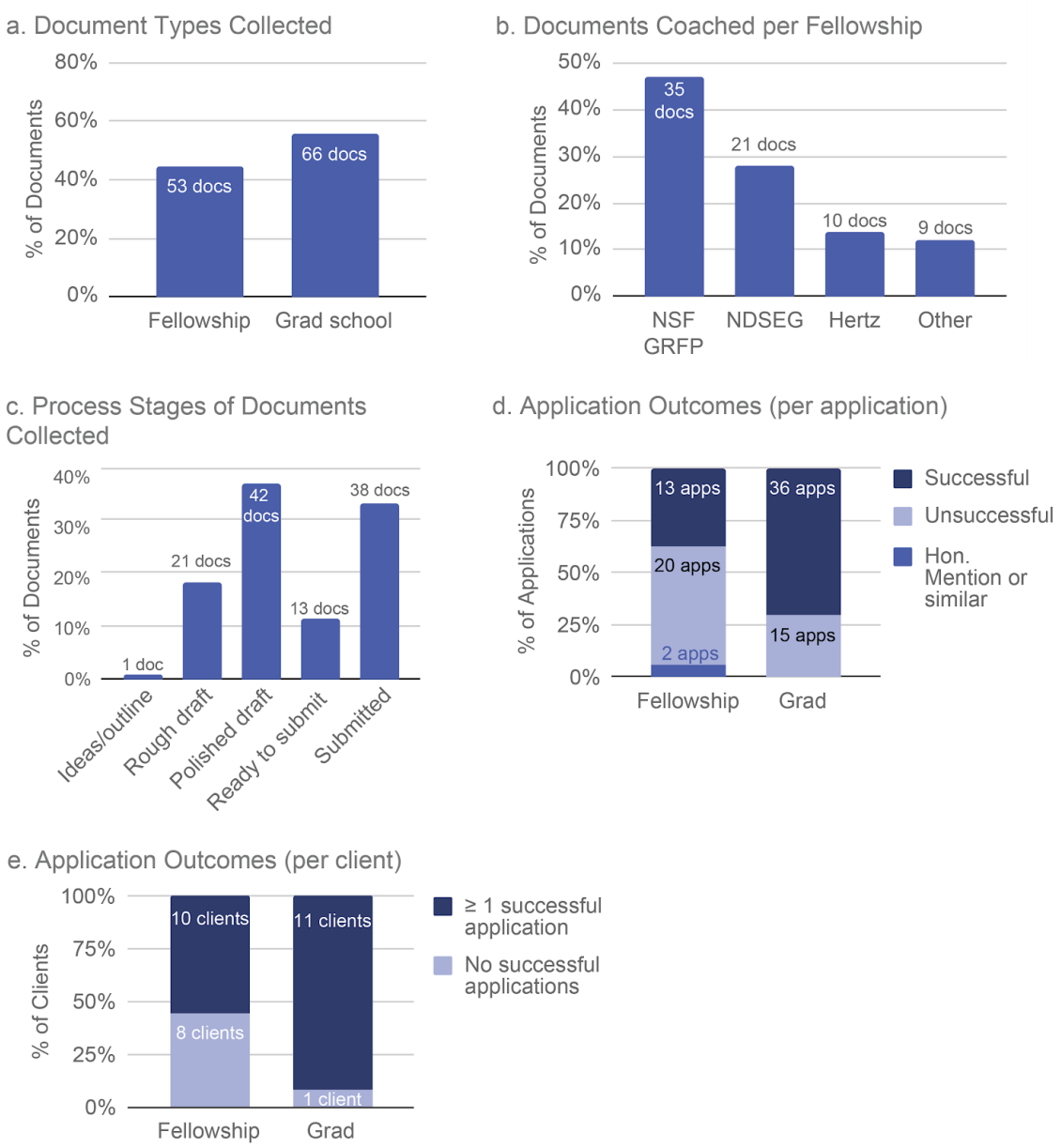


Figure 8: Overview of all documents collected, including **a)** application type: graduate school application vs. fellowship; **b)** type of fellowship, where "other" fellowships include 1-2 each of GEM PhD Engineering Fellowship, NASA NSTRF, Ford Foundation Diversity Fellowship, Paul and Daisy Soros Fellowship for New Americans, National Collegiate Athletic Association Postgraduate Scholarship, and Knight Hennessy Scholarship; and **c**) document's stage of completion (missing data for 4 documents due to incomplete client surveys). Application outcomes are categorized by application **(d)** and by client **(e)**. Bar labels are absolute counts.

Conclusion

In conclusion, we have developed an analytic rubric and a maximum likelihood estimation analysis to quantitatively study the effectiveness of a discipline-specific communication service. These methods were used to analyze a total of 119 personal statement drafts from clients applying for graduate schools and fellowships who voluntarily made appointments with peer coaches from the MIT Communication Lab. The right-skewed

distribution of overall document scores (median = 15.8/21) and largely unchanged individual category scores in previous-next draft comparisons (88%) may be accounted for by the submission of primarily late-stage drafts from a high achieving population. Nevertheless, pairing these statistical approaches with qualitative survey responses was able to provide credible insights.

Altogether, our findings suggest that the Comm Lab's peer coaches may be most impactful in supporting the high-level skills of organization/flow, strategic alignment, and evidence. Moreover, draft comparisons for which the coach and/or client identified major takeaways in a specific category saw universal improvement in that category. This improvement may be a combined result of the clients agreeing upon a clear, actionable list of takeaways during the appointment and/or thinking critically to identify takeaways during completion of the post-appointment survey. To address both of these components, we recommend that peer coaches encourage their clients to generate a list of takeaways at the end of an appointment, and supplement this list with missing takeaways when necessary.

Since organization/flow is both the most mentioned skill in appointment takeaways and a frequently improved rubric category in this study, future Comm Lab efforts may focus on further supporting peer coaches in this area. For example, training for coaches could incorporate more exercises where coaches support clients in reorganizing a draft, and coaches could develop visual schematics to quickly illustrate different organizational options to clients.

Future analyses building on these data and analytical tools could include the following: 1) investigating the large fraction of rubric category scores statistically identified as not having changed, 2) sampling a population with lower self-assessed writing skills and earlier-stage drafts, 3) completing a study on a negative control group of Comm Lab non-users, and 4) comparing our coaches' scores for this dataset with assessments by faculty.

In this work, we have designed and applied an analytical framework to assess the impact of the MIT School of Engineering Communication Lab on the quality of clients' personal statements graduate school and fellowship application; our results suggest that our services are most impactful when coaches and clients agree upon an actionable plan for document iteration, especially when suggestions for improvement focused on high-level skills. Through this work, we highlight several considerations for optimizing experimental design and analytical pipeline for quantitative writing assessment (e.g., post-processing of rubric scores to control for inter-reviewer variability), which can be translated to a range of writing assessments beyond our Comm Lab model, including coursework and more traditional writing centers.

Acknowledgements

For invaluable support of data collection, we gratefully acknowledge the managers and Fellows of the Biological Engineering Communication Lab (especially Dr. Prerna Bhargava for early study design input), Chemical Engineering Communication Lab, Electrical Engineering & Computer Science Communication Lab, Mechanical Engineering Communication Lab. For helpful suggestions about study design, we acknowledge Dr. Anne Marshall (MIT Teaching and Learning Laboratory) and Professor Neal Lerner (Northeastern University).

References

[1] P. Sageev and C. Romanowski, "A Message from Recent Engineering Graduates in the Workplace: Results of a Survey on Technical Communication Skills," *Journal of Engineering Education*, vol. 90, no. 4, pp. 685–693, 2001.

[2] H. J. Passow, “Which ABET Competencies Do Engineering Graduates Find Most Important in their Work?,” *Journal of Engineering Education*, vol. 101, no. 1, pp. 95–118, Jan. 2012, doi: 10.1002/j.2168-9830.2012.tb00043.x.
[3] N. A. Bousaba, J. M. Conrad, J. L. Coco, S. M. Miri, and R. W. Cox, “Improving oral presentation in an electrical and computer engineering department,” in *Proceedings of the 2014 ASEE Annual Conference*, Indianapolis, IN, 2014.
[4] P. Hager and S. Holland, *Graduate Attributes, Learning and Employability*, 1st ed., vol. 6. Springer Netherlands, 2006.
[5] D. Jackson, “An international profile of industry-relevant competencies and skill gaps in modern graduates,” *The International Journal of Management Education*, vol. 8, pp. 29–58, Apr. 2010.
[6] J. A. Donnell, “Why Industry Says That Engineering Graduates Have Poor Communication Skills: What the Literature Says,” p. 13.
[7] P. M. Berthouex, “Honing the Writing Skills of Engineers,” *Journal of Professional Issues in Engineering Education and Practice*, vol. 122, no. 3, pp. 107–110, Jul. 1996, doi: 10.1061/(ASCE)1052-3928(1996)122:3(107).
[8] Institute-wide Task Force on the Future of MIT Education, “Final Report,” 2014.
[9] A. Hanson, P. Lindahl, S. Strasser, A. Takemura, D. Englund, and J. Goldstein, “Technical Communication Instruction for Graduate Students: The Communication Lab vs, A Course,” in *2017 ASEE Annual Conference & Exposition Proceedings*, Columbus, Ohio, 2017, p. 28932, doi: 10.18260/1-2--28932.
[10] B. Bloom, “The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring,” *Educational Researcher*, vol. 13, no. 6, pp. 4–16, 1984.
[11] K. J. Topping, “Trends in Peer Learning,” *Educational Psychology*, vol. 25, no. 6, pp. 631–645, Dec. 2005, doi: 10.1080/01443410500345172.
[12] J. Casey, “The Relationship Between Writing Centers and Improvement in Writing Ability: An Assessment of the Literature,” *Education*, vol. 122, no. 1, 2001.
[13] C. Bazerman, *Reference guide to writing across the curriculum*. West Lafayette, Ind.: Parlor Press : WAC Clearinghouse, 2005.
[14] British Computer Society, “Guidelines on Course Accreditation,” 2015.
[15] “Criteria for Accrediting Engineering Programs, 2016 – 2017 | ABET.” [Online]. Available: http://www.abet.org/accreditation/accreditation-criteria/criteria-for-accrediting-engineering-programs-2016-2017/. [Accessed: 13-Jul-2016].
[16] K. B. Vasel, “The Skills Employers Wish College Grads Had,” *FOXBusiness*, 30-Jan-2014. [Online]. Available: http://www.foxbusiness.com/features/2014/01/30/skills-employers-wish-college-grads-had.html. [Accessed: 12-Jun-2016].
[17] McCuen@aacu.org, “Falling Short? College Learning and Career Success,” *Association of American Colleges & Universities*, 26-Jan-2015. [Online]. Available: https://www.aacu.org/leap/public-opinion-research/2015-survey-falling-short. [Accessed: 13-Jul-2016].
[18] D. Belkin, “Test Finds College Graduates Lack Skills for White-Collar Jobs,” *Wall Street Journal*, 16-Jan-2015.
[19] J. Y. Yoritomo *et al.*, “Examining engineering writing instruction at a large research university through the lens of writing studies,” p. 22.
[20] D. V. Svihla, “Peer Review and Reflection in Engineering Labs: Writing to Learn and Learning to Write,” p. 25.
[21] M. M. Alley, P. S. University, and U. Park, “Using Undergraduate Mentors to Scale the Teaching of Engineering Writing,” p. 17.
[22] D. S. Summers, “Experiments in the Communication Lab: Adaptations of the Comm Lab Model in Three Institutions,” p. 20.
[23] S. Dinitz and S. Harrington, “The role of disciplinary expertise in shaping writing tutorials,” *The Writing Center Journal*, pp. 73–98, 2014.
[24] R. Weissbach and R. Pflueger, “Technical writing knowledge transfer from first year composition to major courses,” in *Proceedings of the 2014 IEEE Frontiers in Education Conference (FIE)*, 2014.
[25] L. R. Hughes, “Tutoring technical documents in the writing center: implications for tutor training and practices,” Texas Tech University, 2009.
[26] J. Mackiewicz, “The Effects of Tutor Expertise in Engineering Writing: A Linguistic Analysis of

Writing Tutors' Comments," *IEEE Transactions on Professional Communication*, vol. 47, no. 4, pp. 316–328, Dec. 2004, doi: 10.1109/TPC.2004.840485.
[27] J. D. Bransford, A. L. Brown, and R. R. Cocking, *How people learn*. Washington, DC: National Academy Press, 2000.
[28] F. M. Newmann, H. M. Marks, and A. Gamoran, "Authentic pedagogy and student performance," *American Journal of Education*, pp. 280–312, 1996.
[29] M. M. Lombardi, "Authentic learning for the 21st century: An overview," *Educause learning initiative*, vol. 1, no. 2007, pp. 1–12, 2007.
[30] R. Day Babcock and T. Thonus, "A sample research question: What is a successful tutorial?," in *Researching the Writing Center*, New York: Peter Lang, 2012, pp. 143–169.
[31] K. A. Bruffee, "Collaborative Learning and the 'Conversation of Mankind,'" *College English*, vol. 46, no. 7, p. 635, Nov. 1984, doi: 10.2307/376924.
[32] D. Eubanks, "A Guide for the Perplexed," *Intersection (Association for the Assessment of Learning in Higher Education)*, no. Fall 2017, pp. 4–13, 2017.
[33] N. Lerner, "Choosing beans wisely," *The Writing Lab Newsletter*, vol. 26, no. 1, 2001.
[34] *Teaching and Assessing Writing*. .
[35] R. C. Calfee and R. G. Miller, "Best Practices in Writing Assessment," p. 13.
[36] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data Via the *EM* Algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, Sep. 1977, doi: 10.1111/j.2517-6161.1977.tb01600.x.
[37] L. Salem, "Decisions … Decisions: Who Chooses to Use the Writing Center?," *Writing Center Journal*, vol. 35, no. 2, p. 147, 2016.
[38] J. C. Hu, "NSF graduate fellowships disproportionately go to students at a few top schools," *Science Magazine*, 26-Aug-2019. [Online]. Available: https://www.sciencemag.org/careers/2019/08/nsf-graduate-fellowships-disproportionately-go-students-few-top-schools. [Accessed: 02-Feb-2020].

Appendix: Rubric and surveys

		Absent/Incomplete (0)	Not Competitive (1)	Somewhat Competitive (2)	Highly Competitive (3)
	Visual Impact	Formatting elements detract from the visual appeal of the document. Formatting is such that document cannot be skimmed.	Formatting elements somewhat detract from the visual appeal of the document. Formatting makes it difficult for the reader to skim the document.	Formatting elements neither add to nor detract from the visual appeal of the document. Formatting does not detract from the reader's ability to skim the document, e.g., no headings, bolding, or other forms of emphasis/distinction.	Formatting elements enhance the visual appeal of the document. Formatting elements (such as headings, bolding, or other forms of emphasis/distinction) enhance the reader's ability to skim the document.
Why	Strategic Alignment	Both future goals and the benefits of this award are absent. Document uses neither intellectual merit nor broader impacts.	Document vaguely discusses candidate's future goals but does not relate them to NSF's mission or this award and/or includes substantial extraneous material. Document uses only intellectual merit or broader impacts.	Document concretely discusses candidate's future goals but does not explain how the GRFP would support those goals and/or includes extraneous material that slightly distracts the reader. Document heavily relies on either intellectual merit or broader impacts.	Document concretely discusses candidate's future goals, their alignment with the organization's mission, and how the desired outcome would support those goals with little to no extraneous information. Document clearly explains the value of the candidate in terms of both intellectual merit and broader impacts (for NSF GRFP) or through a combination of academic, research, and extracurricular activities (for grad school applications).
Who	Audience	Document is incomprehensible to reader outside of candidate's specific field. Document contains many instances of jargon with no explanation of terms.	Document is difficult for reader outside of candidate's specific field to understand or is too high level to convey expertise. Document contains a noticeable amount of unnecessary jargon or unexplained terms.	Document has parts that are difficult to understand or is somewhat too high level. Document contains a small amount of unnecessary jargon or unexplained terms.	Document is easily comprehensible to the educated non-expert and conveys candidate's expertise. Document contains minimal instances of jargon, and these terms are explained.
What	Context	Research discussion is either entirely background or missing any background. Connection of research and/or experiences with broader impacts is completely absent.	Background of research is either highly dominant or largely absent in the document. Connection of research and/or experiences with broader impacts is generally unclear.	Background of research is either unnecessary or insufficient to place candidate's work in the context of the field. Connection of research and/or experiences with broader impacts are occasionally unclear.	Background of research is both necessary and sufficient to place candidate's work in the appropriate context of the field Document communicates the connection of the candidate's research and/or experiences with broader impacts.
	Evidence	Arguments/evidence for the candidate's abilities are absent. Document suggests candidate does not have the potential to become a successful researcher.	Arguments/evidence for the candidate's abiliites are weak and not based in concrete examples of past successes. Document does not convey confidence in the candidate's potential to become a successful researcher.	Arguments/evidence for the candidate's abilities are somewhat convincing with limited examples of past successes. Document conveys some confidence in the candidate's potential to become a successful researcher.	Arguments/evidence for the candidate's abilities are compelling with several concrete examples of past successes. Document conveys confidence in the candidate's potential to become a successful researcher.
How	Organization/Flow	Senteces/paragraphs are not logically connected; there is no comprehensible organization. Transitions are absent.	Senteces/paragraphs are weakly connected with some comprehensible organization. Transitions are not clear or effective.	Senteces/paragraphs are largely connected and organized, with few noticeable exceptions. Transitions are clear but do not effectively segue into next argument.	Sentences/paragraphs are connected and organized throughout. Transitions are clear and effectively segue into next argument.
	Language Mechanics	Reader is distracted by several instances of misspelling or incorrect grammar. Language makes it difficult to understand content or purpose.	Reader notices more than one instance of misspelling or incorrect grammar. Language is overly wordy but comprehensible.	Reader notices an instance of misspelling or incorrect grammar. Language is mostly concise, concrete, and active.	Grammar and spelling are not distracting to the reader. Language is concise, concrete, and active.

Rubric Instructions

Please evaluate the NSF personal statement or graduate school personal statement from the perspective of an [Institution] Communication Lab coach. There is no need to imagine that you are an NSF reviewer or graduate admissions committee member. With the coach mindset, follow the instructions below in order:

1. Read the full rubric in detail before beginning a scoring session.

2. Skim the document for up to 30 seconds and then give scores to the prompts in the Visual Impact category of the rubric.

3. Now read the document in full. This should take 5 – 10 minutes.

4. Fill out the rubric in your **personal** scoring spreadsheet (Dropbox: Comm Lab – Ed Studies >Data Analysis > Data > Drafts by scorer > [Your folder]), referring back to the document as necessary.
    a. Numerically evaluate every prompt.
    b. Default to giving whole numbers. You may use half-points if necessary.
    c. When calculating the total for each rubric row, average the subscores. However, avoid assigning quarter points; use your gut feeling to round up or down to X.0 or X.5.

5. If a draft has incomplete or missing content…

    a. Mark the "INCOMPLETE" checkbox in your scoring spreadsheet.
    b. If a category of content (e.g., match) is **completely missing**, without a placeholder indicating awareness that content should be there, add 0 points for the missing content.
    c. If content is **completely missing** but there is a **placeholder** ("Insert paragraph about professors I'd want to work with here"), add 0.5 points for the missing content.
    d. If content is **partial** and there is a **placeholder**: Score whatever is there leniently (under the assumption that later additions might synergize to improve).

6. In the "Comments & Impressions" column of the spreadsheet, please record a ~1-sentence summary of your overall impression of the draft. E.g., "This essay had an engaging story line and the candidate seemed excited and committed to a PhD despite lack of research experience."
    - In particular, note down any strong biases or gut reactions that you had about the content. We know that these tend to bias the numerical scores.

Instructions continue below

7. **Calibration:** When you are **1/3** and **2/3** of the way through your assigned drafts, please re-calibrate your scoring by using the following two drafts as calibration points. They were selected based on very close score agreement among at least three scorers:

		Scores assigned in "Summer 2019 Scoring:"			
		Reviewer 1	Reviewer 2	Reviewer 3	Reviewer 4
Low calibration point	Gradschool/f0200bf6572ZUBavpR9yNfUc8.docx	12	12.5	14	11.5
		Reviewer 1	Reviewer 2	Reviewer 3	Reviewer 4
High calibration point	Fellowship/81ae1b0cb5uyyJgYtiohaGFvb.pdf	18	19	20	(did not score since reviewer coached this client)

Calibrate as follows:

a. Re-read the two calibration points and take note of the scores previously assigned.
b. Review the scores that you assigned to the last two drafts that you reviewed. Are the scores generally consistent with how the calibration points were scored?
c. If not, think about what might have shifted in your scoring habits or mindset. Try to return to the mindset you used when the calibration points were scored.
d. Feel free to note down any observations about the calibration process in your scoring spreadsheet's comments area.

Rubric Glossary of Terms

Active language	Using primarily active voice and specific action verbs *e.g. Investigated, analyzed, characterized, developed*
Audience	Educated non-experts who evaluate the document according to NSF's guidelines
Benefit of award	How receiving the fellowship or admission into the graduate program will help the candidate achieve their goals
Broader impacts	The potential to benefit society and contribute to the achievement of specific, desired societal outcomes *e.g. Impact of research itself, dissemination of results, involvement in activities related to teaching, mentoring, and increasing participation of underrepresented groups*
Context	The introduction to each topic which communicates relevance and importance
Evidence	Specific examples that highlight the candidate's experiences and potential
Formatting elements	White space, margins, indentations, headings, bold, italics, underline, etc. *e.g. Section headers of broader impacts and intellectual merit*
Future goals	The candidate's projected career path and motivation for pursuing a PhD
Intellectual merit	The potential to advance knowledge *e.g. Description of previous research experience, publications, presentations, and future goals*
Jargon	Any vocabulary that may be unfamiliar to an educated non-expert *e.g. Small-angle X-ray scattering (SAXS) is a technique used to characterize materials on the nano-scale.*
Language mechanics	Writing conventions including spelling, grammar, and word choice
NSF's mission	The purpose of the NSF GRFP is to help ensure the vitality and diversity of the scientific and engineering workforce of the United States.
Organization/Flow	The structure of the document including sections, paragraphs, and sentences, and the connections between each structural element

Strategic alignment	The ability of the document to address the review criteria and support the organization’s mission
Transitions	Sentences that connect other sentences, paragraphs, or sections
Visual impact	The appeal of the document based on skimming for ~ 30 seconds

Post-Appointment Survey for Peer Coaches

[Institution] Communication Lab Study

Post-appointment reflections

All data will be anonymized prior to analysis.

Any personally identifying information requested below will be used only to match your survey response with your client's.

If this is your first time submitting a post-appointment survey, please make sure you send in your consent form to [contact].

What is the email associated with **your** Comm Lab account?

What is the email of the **client** regarding whom you are filling out this survey?

On what date did this appointment take place?

←I December 2018 →I

Su	Mo	Tu	We	Th	Fr	Sa
25	26	27	28	29	30	1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29

←I December 2018 →I

Su	Mo	Tu	We	Th	Fr	Sa
30	31	1	2	3	4	5

How long was this appointment?

- ○ 30 minutes
- ○ 60 minutes

Please answer the questions below with respect to your discussions about the client's personal statement(s).

If you discussed documents other than personal statements, disregard those portions of the discussion for these questions.

What did you feel was the biggest takeaway for the client (e.g., most important skill discussed) from this coaching appointment?

To what extent did you and your client address the following areas during your appointment?

	Not at all	A little	A lot
Aligning their document with their **goals** E.g., making a clear argument that they are a good match for a fellowship/grad school	○	○	○
Appropriately addressing their **audience** E.g., conveying technical expertise without using confusing jargon	○	○	○
Including appropriate **context and motivation** E.g., explaining why their past research work is interesting	○	○	○
Including persuasive **ideas/evidence** E.g., concrete examples that demonstrate their research or leadership credentials	○	○	○
Creating effective **organization/flow** E.g., creating transitions between paragraphs	○	○	○
Using appropriate **language/grammar** E.g., getting rid of typos or clarifying confusing sentences	○	○	○

How similar was your client’s technical expertise to your own?

Not at all close Fundamental concepts of their field were not initially clear to me.	A little close	Somewhat close I was familiar with some concepts of their field, a lot was not initially clear to me.	Close	Nearly identical We might as well be in the same lab.	N/A Couldn’t tell
○	○	○	○	○	○

To what extent did you use technical knowledge of your client's scientific/engineering field to help them during the appointment?

E.g., helping them...

- think about what critiques a fellowship judge from a given field would be likely to have
- explain the significance of a research project to faculty members from a given field
- by suggesting, “I think a professor in machine learning would wonder about…”

Not at all No technical content was addressed; they could have had a similar discussion with someone who wasn’t a scientist/engineer.	A little Technical content was addressed only occasionally.	Moderately Our discussion relied on technical understanding, but not background from their specific field.	A lot They couldn’t have had a similar discussion with someone inexperienced in their specific field.
○	○	○	○

Post-Appointment Survey for Clients

[Institution] Communication Lab Study

Identity

Thank you for participating in the Comm Lab study!

This survey will ask you to upload the **personal statement draft(s)** you discussed in your appointment. If you don't have access to the relevant files on your current device, as they stood before the appointment, please return to the survey from a device that has access to these files. (Uploading files from a mobile device with access to cloud services such as Dropbox may work, but it is not always reliable.)

- [ ] Yes, at least part of my appointment concerned a personal statement.
- [ ] Yes, I have access to my personal statement draft(s).

What is the email associated with your Communication Lab profile?

Have you previously filled out a post-appointment survey for this Communication Lab study?

"Yes" means you've previously filled out a survey page where you specified your MIT affiliation, your English proficiency, and your skill level in written communication.

- ○ Yes
- ○ No

Intake

Please identify yourself as one of the following.

- ○ Undergraduate
- ○ Graduate student, year 1
- ○ Graduate student, year 2
- ○ Graduate student, year 3
- ○ Graduate student, year 4+
- ○ Research technician or other lab staff
- ○ [ ] Other

If English is NOT your first/home language, how would you rate your fluency in written English?

Very poor	Poor	Fair	Good	Excellent	N/A
○	○	○	○	○	○

How strong do you consider your abilities in written communication to be, independent of language?

Very poor	Poor	Fair	Good	Excellent
○	○	○	○	○

Appointment info

All data about your appointments and drafts will be anonymized prior to analysis and stored in a secure fashion.

With which Communication Lab was the appointment for which you're submitting this survey?

- ◯ Biological Engineering
- ◯ Chemical Engineering
- ◯ Electrical Engineering & Computer Science
- ◯ Mechanical Engineering

On what date did this appointment take place?

←I December 2018 →I

Su	Mo	Tu	We	Th	Fr	Sa
25	26	27	28	29	30	1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31	1	2	3	4	5

Which type of document did you work on during this appointment?

- ◯ Fellowship application - personal statement
- ◯ Grad school application - personal statement
- ◯ Both

Fellowship personal statement

Please answer the questions below with respect to your fellowship personal statement. If you worked on a grad school statement, too, we'll ask you about that on the next page.

Are you going to use this fellowship personal statement draft as the basis for multiple applications (i.e., you plan to tailor it later for specific fellowships), or are you preparing a draft for one specific application?

- ○ Using for multiple applications
- ○ Using for one specific application

What is the first date when you will need to submit this document?

←I December 2018 →I

Su	Mo	Tu	We	Th	Fr	Sa
25	26	27	28	29	30	1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31	1	2	3	4	5

Which specific fellowship(s) do you intend to use this personal statement for?

Select all that apply.

- [ ] NSF GRFP
- [ ] NDSEG
- [ ] Hertz
- [ ] Ford Foundation Diversity Fellowship
- [ ] DOE CSGF (computational science)
- [ ] Other
- [ ] Other
- [ ] Other

Since your previous appointment with the Comm Lab (if any), have you received support/feedback on this personal statement from resources other than the Comm Lab? If so, please estimate below how many hours of support/feedback you have received.

Lab advisor or other faculty	
Undergraduate colleagues	
Graduate student or postdoc colleagues	
Other friends/family	
MIT Writing and Communication Center	
MIT Career Advising and Professional Development	
Other	
Other	
Other	

Please upload the personal statement draft you discussed during this appointment.

If you didn't have a written draft yet (e.g., you were still brainstorming), please upload a blank Word document named “blank.doc” or “blank.docx” here.

As of the time of the appointment, how ready was this personal statement to be used in an application?

Documents at all stages of readiness are welcome; we just want to get a sense of how you view the state of your document.

I was still brainstorming	Ideas/outline, but no draft I knew what I wanted to say, but it needed to be turned into prose	Rough draft Content was mostly there, but it needed a lot of work	Polished draft Still needed a bit of work, but the draft was nearly there	Draft ready I'd have been willing to send it out as-is; just needed final checks
○	○	○	○	○

How much do you think your document has improved or will improve as a result of this appointment?

Not at all improved	Slightly improved	Moderately improved	Very improved	Extremely improved
○	○	○	○	○

Graduate school personal statement

Please answer the questions below with respect to your graduate school personal statement.

Are you going to use this graduate school personal statement draft as the basis for multiple applications (i.e., you plan to tailor it later for specific schools), or are you preparing a draft for one specific application?

- ◯ Using for multiple applications
- ◯ Using for one specific application

What is the first date when you will need to submit this document?

←I December 2018 →I

Su	Mo	Tu	We	Th	Fr	Sa
25	26	27	28	29	30	1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31	1	2	3	4	5

Which specific graduate school program(s) do you intend to use this personal statement for?

Please list one school per line.

Since your previous appointment with the Comm Lab (if any), have you received support/feedback on this personal statement from resources other than the Comm Lab? If so, please estimate below how many hours of support/feedback you have received.

Lab advisor or other faculty	
Undergraduate colleagues	
Graduate student or postdoc colleagues	
Other friends/family	
MIT Writing and Communication Center	
MIT Career Advising and Professional Development	
Other	
Other	
Other	

Please upload the personal statement draft you discussed during this appointment.

If you didn't have a written draft yet (e.g., you were still brainstorming), please upload a blank Word document named “blank.doc” or “blank.docx” here.

As of the time of the appointment, how ready was this personal statement to be used in an application?

Documents at all stages of readiness are welcome; we just want to get a sense of how you view the state of your document.

I was still brainstorming	Ideas/outline, but no draft I knew what I wanted to say, but it needed to be turned into prose	Rough draft Content was mostly there, but it needed a lot of work	Polished draft Still needed a bit of work, but the draft was nearly there	Draft ready I'd have been willing to send it out as-is; just needed final checks
○	○	○	○	○

How much do you think your document has improved or will improve as a result of this appointment?

Not at all improved	Slightly improved	Moderately improved	Very improved	Extremely improved
○	○	○	○	○

Post-appointment reflections

Please answer the questions below with respect to your discussions about your personal statement(s).

If you discussed documents other than personal statements, disregard those portions of the discussion for these questions.

What was your biggest takeaway (e.g., most important skill discussed) from this coaching appointment?

To what extent did you and your Communication Fellow/Advisor address the following areas during your appointment?

	Not at all	A little	A lot
Aligning your document with your **goals** E.g., making a clear argument that you are a good match for a fellowship/grad school	○	○	○
Appropriately addressing your **audience** E.g., conveying technical expertise without using confusing jargon	○	○	○
Including appropriate **context and motivation** E.g., explaining why your past research work is interesting	○	○	○
Including persuasive **ideas/evidence** E.g., concrete examples that demonstrate your research or leadership credentials	○	○	○
Creating effective **organization/flow** E.g., creating transitions between paragraphs	○	○	○
Using appropriate **language/grammar** E.g., getting rid of typos or clarifying confusing sentences	○	○	○

Please answer the questions below with respect to your discussions about your personal statement(s).

If you discussed documents other than personal statements, disregard those portions of the discussion for these questions.

How helpful do you think skills/ideas discussed during this appointment would be when creating other personal statements in the future?

Not at all helpful	A little helpful	Moderately helpful	Very helpful	Extremely helpful
○	○	○	○	○

How helpful do you think skills/ideas discussed during this appointment would be when creating other types of technical communication in the future, such as an abstract, grant proposal, or presentation?

Not at all helpful	A little helpful	Moderately helpful	Very helpful	Extremely helpful
○	○	○	○	○

How similar was your Communication Fellow/Advisor's technical expertise to your own?

Not at all close I had to explain fundamental concepts of my field to them	A little close	Somewhat close They were familiar with some concepts of my field, but I had to explain a lot	Close	Nearly identical We might as well be in the same lab	N/A Couldn't tell
○	○	○	○	○	○

To what extent did your Communication Fellow use technical knowledge of your scientific/engineering field to help you during your appointment?

E.g., helping you...

- think about what critiques a fellowship judge from a given field would be likely to have
- explain the significance of a research project to faculty members from a given field
- by suggesting, "I think a professor in machine learning would wonder about…"

Not at all No technical content was addressed; I could have had a similar discussion with someone who wasn't a scientist/engineer.	A little Technical content was addressed only occasionally.	Moderately Our discussion relied on technical understanding, but not background from my specific field.	A lot I couldn't have had a similar discussion with someone inexperienced in my specific field.
○	○	○	○