Gauging Engagement: Measuring Student Response to a Large-Scale College Advising Field Experiment

Brian Heseung Kim
University of Virginia

Katharine Meyer
Brown University

Alice Choe
University of Virginia

Interactive, text message-based advising programs have become an increasingly common strategy to support college access and success for underrepresented student populations. Despite the proliferation of these programs, we know relatively little about how students engage in these text-based advising opportunities and whether that relates to stronger student outcomes – factors that could help explain why we've seen relatively mixed evidence about their efficacy to date. In this paper, we use data from a large-scale, two-way text advising experiment focused on improving college completion to explore variation in student engagement using nuanced interaction metrics and automated text analysis techniques (i.e., natural language processing). We then explore whether student engagement patterns are associated with key outcomes including persistence, GPA, credit accumulation, and degree completion. Our results reveal substantial variation in engagement measures across students, indicating the importance of analyzing engagement as a multi-dimensional construct. We moreover find that many of these nuanced engagement measures have strong correlations with student outcomes, even after controlling for student baseline characteristics and academic performance. Especially as virtual advising interventions proliferate across higher education institutions, we show the value of applying a more codified, comprehensive lens for examining student engagement in these programs and chart a path to potentially improving the efficacy of these programs in the future.

VERSION: April 2022

Suggested citation: Kim, Brian Heseung, Katharine Meyer, and Alice Choe. (2022). Gauging Engagement: Measuring Student Response to a Large-Scale College Advising Field Experiment. (EdWorkingPaper: 22-567). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/dcrq-rv92

Gauging Engagement: Measuring Student Response to a Large-Scale College Advising Field Experiment

Brian Heseung Kim
University of Virginia
bhk5fs@virginia.edu

Katharine Meyer
Brown University
katharine_meyer@brown.edu

Alice Choe
University of Virginia
ac2mj@virginia.edu

Draft as of March 15, 2022[1]

Abstract

Interactive, text message-based advising programs have become an increasingly common strategy to support college access and success for underrepresented student populations. Despite the proliferation of these programs, we know relatively little about *how* students engage in these text-based advising opportunities and whether that relates to stronger student outcomes – factors that could help explain why we've seen relatively mixed evidence about their efficacy to date. In this paper, we use data from a large-scale, two-way text advising experiment focused on improving college completion to explore variation in student engagement using nuanced interaction metrics and automated text analysis techniques (i.e., natural language processing). We then explore whether student engagement patterns are associated with key outcomes including persistence, GPA, credit accumulation, and degree completion. Our results reveal substantial variation in engagement measures across students, indicating the importance of analyzing engagement as a multi-dimensional construct. We moreover find that many of these nuanced engagement measures have strong correlations with student outcomes, even after controlling for student baseline characteristics and academic performance. Especially as virtual advising interventions proliferate across higher education institutions, we show the value of applying a more codified, comprehensive lens for examining student engagement in these programs and chart a path to potentially improving the efficacy of these programs in the future.

[1] We are thankful for the fantastic research support of Danielle Peacock and Emily Arizaga. The Nudges to the Finish Line (N2FL) intervention was generously supported by the U.S. Department of Education (Institute of Education Sciences) under the grant number R305N160026. Research work was also supported by the Institute of Education Sciences Virginia Education Science Training Pre-Doctoral Fellowship under grant number R305B140026.

I. Introduction

Despite the high economic returns to college completion (Avery & Turner, 2012; Carnevale, Jayasundera, & Gulish, 2016), just over half of students who enroll at college have attained a bachelor's degree (Bound, Lovenheim, & Turner, 2010; Denning, Eide, & Warnick, 2019; Shapiro et al., 2016). Colleges and non-profits have invested in various strategies to improve college completion, ranging from resource-intensive advising (Scrivener et al., 2015) to light-touch messaging campaigns (Castleman & Page, 2016). More recently, organizations have implemented hybrid text-based advising models that provide light-touch supports *as well as* personalized advising to students after matriculation (Gurantz et al., 2020; Oreopoulos & Petronijevic, 2018; Page & Gehlbach, 2017; Sullivan et al., 2019).

Given that we have increasing evidence for the limited effectiveness of *one*-way texting campaigns (a la many nudging interventions) when scaled to broader contexts and populations (Bird et al., 2021), we might expect the majority of benefits from these text-based advising campaigns to then come from the *two*-way interactions between students and their advisors. Despite this intuition, past evaluations have found that the effectiveness of two-way text-based advising in improving college persistence and graduation *also* varies across contexts and students (Avery et al., 2021). It remains unclear exactly what factors contribute to this variation in program effectiveness, but one potential explanation that remains underexplored is the variability in student engagement both within and across interventions. If the two-way interactions between students and advisors are the key mechanism for program effectiveness, the extent of student engagement in any text-based advising program can be thought of as a form of endogenous intensity for the intervention, where some students choose to engage in and "receive" the intervention at greater intensities than others.

While some scholars have begun attempting to document variation in student engagement in such text-based advising campaigns (e.g., Fesler, 2020; Arnold et al., 2020), how best to operationalize and measure student engagement remains uncodified, and the relationship of any such engagement with student outcomes is still ambiguous. And though there exist some consistent measures of engagement used in interactive text campaigns in other contexts (e.g., smoking cessation texting programs), these tend to be simplistic and overly focused on response rates. A deeper understanding of appreciable differences in student behaviors can help scholars generate and test hypotheses around which behaviors are potentially malleable through an intervention and would lead to improved student outcomes. For example, if we observe a strong positive correlation between academic persistence and students who more frequently solicit assistance from their advisors, this insight could motivate future research into the causality of this relationship through thoughtful experimental design (i.e., conducting an interactive messaging experiment where one treatment wing more explicitly solicits student requests for assistance in its prompts). Our descriptive exploration is then intended to highlight future venues for research that could, eventually, refine the design of two-way text advising programs and improve their efficacy. Such design insight has become increasingly important to gather as an increasing number of higher education institutions have turned to remote advising practices and campus-wide text messaging campaigns to support their students.

In this study, we seek to refine our understanding of student engagement in text-based advising using a variety of data- and text-mining (i.e., natural language processing, or NLP) techniques to analyze student-advisor interactions from the Nudges to the Finish Line (N2FL) text-advising intervention (Bettinger, et al., 2021). N2FL was randomly delivered to students approaching degree completion at over twenty colleges and universities in five states across the U.S.

Treated students in the intervention received pre-scheduled and pre-written messages (what we herein refer to as "scheduled messages") that provided information about important deadlines and encouraged use of campus resources like academic tutoring centers and financial aid. Because N2FL was designed as a two-way interactive campaign, the scheduled messages encouraged students to write back and engage in impromptu conversations with advisors at their campus (e.g., "...Do you need help with applying for financial aid?").

To quantify variation in how students engaged with the intervention, we employ a wide array of measures related to the intensity, duration, response speed, and content of student-written messages (what we herein refer to as "student replies"). For example, we examine the proportion of scheduled messages students responded to, the number of requests for help students made, the positivity or negativity of student texts (known as "sentiment analysis" in NLP), and the extent to which students discussed various topics of conversation (known as "topic modeling" in NLP). We rely on approaches from prior literature to form these measures, while also constructing several novel measures to explore. In general, we find wide variation in nearly every measure we construct, emphasizing the variation in students' interaction with text-based advising interventions. We moreover interpret measures with wide variation as more likely to be malleable (i.e., when compared with measures that have extremely narrow distributions across students), though we are unable to explore this explicitly in this study. We also find that this array of engagement measures tend to be uncorrelated with one another, indicating that patterns and intensities of student engagement are multi-faceted and are unlikely to be well-captured using simple response rates as is the current business-as-usual in the field.

After describing the construction of these measures and examining the extent of variation in student engagement along these measures in N2FL, we proceed to descriptively examine how these measures correlate with the outcomes of interest in the intervention: re-enrollment, credit accumulation, GPA, and degree receipt in each term following the start of the intervention. We find that even after controlling for student baseline characteristics and academic performance, there exist large and persistent relationships between engagement measures and student outcomes. For example, greater frequency of student engagement was strongly related to credit accumulation, GPA, and eventual degree receipt. We also observe higher GPA and credit accumulation among students whose responses were more positive in tone, and more mixed relationships with outcomes based on the topics of discussion students brought up in their replies. While not causal, these results suggest several pathways for future research to more explicitly test the malleability and impact of these more promising engagement measures.

Taken together, our findings offer two main contributions to the field. First, this analysis demonstrates the value of examining text interaction data and student engagement in the context of two-way text-based interventions. We find meaningful variation in engagement behaviors. That engagement varies implies the opportunity to change engagement behaviors, highlighting opportunities to test different strategies to affect student engagement. While this is a descriptive study and student engagement patterns in this context are endogenous, we nonetheless view this as an important initial step in identifying the engagement patterns correlated with college persistence that future interventions might target through careful programmatic design. An increasing number of higher education institutions are turning to text-based advising as a cost-effective tool to reach students at scale. Without a stronger understanding of the potential mechanisms underlying text-based support programs' efficacy, institutions run the risk of designing interventions that do not meaningfully engage students and therefore do not meaningfully improve student outcomes. This paper thus draws important attention to student engagement as a valuable and multi-dimensional mechanism demanding more consistent study in program evaluation and design.

Second, we add to the growing literature showcasing robust text-as-data/NLP methods to enhance our understanding of large-scale educational text data and field experiments. The measures we deploy here are imminently usable across any two-way texting intervention given the common collection of text interaction data via large-scale messaging platforms (e.g., Signal Vine), and we "open-source" our code and methodology to facilitate continued iteration on these engagement measures more broadly. Versions of this framework are also imminently applicable to many two-way educational interactions captured in text, such as discussion board posts, email exchanges, virtual tutoring, and other texting-based interventions. Our hope is to advance the field's ability to apply more comprehensive and codified engagement measures more broadly. Through a combination of simple (e.g., keyword-based) and sophisticated (e.g., neural network-based) NLP methodologies, we are able to provide nuanced insights about how students engage in the two-way advising program – both *at scale* and *in real-time*. While close qualitative reading of the text conversations in this intervention would be instructive in its own right, the scale of analysis and rapidity of insights afforded by these automated text analysis systems (e.g., using a single analyst to examine thousands of text messages almost immediately after their receipt) allow us to gain a far greater understanding of interventions both after the fact and as they happen – unlocking immense potential in program design and evaluation going forward.

In Section II we summarize insights from other studies of engagement in text-based outreach, and in Section III we describe the intervention context and student sample. We outline our methodology in Section IV and share results in Section V. We conclude with a discussion of our results and the implications for future research and practice in Section VI.

II. Prior Literature on Measuring Engagement

Despite a proliferation of text message intervention studies across domains (e.g., education, healthcare, finance, political campaigns), the literature is sparse and disconnected around how to define participant engagement as a construct. Most commonly, studies adopt simplified definitions that rely on aggregate response rates, such as overall responsiveness rate (e.g., "high" is ≥90%, "good" is ≥70–<90%, and "low" is <70%, per Zhang et al., 2018), absolute thresholds for response counts (e.g., 10 or fewer responses, 10-20 responses, or more than 20 responses, per Irvine et al., 2017), and responsiveness by message category (e.g., prompted versus unprompted text messages, per Psihogios et al., 2019).

Other studies have also chosen to define engagement in more contextually-specific ways, further complicating the process of developing a more unified definition. For example, Nelson (2020) initially presented participants with a choice about how frequently they wanted to receive text messages; accordingly, they defined engagement using a joint measure of participants' stated preference for messaging and their ensuing response rates. Another two-way text messaging study designed to reduce binge drinking defined engagement, in part, by manually coding participant responses for whether the informational texts were understood correctly (e.g., participants responded with personal and specific details that demonstrated they cognitively processed how the messages related to their own lives) and whether responses were related to specific components of the behavior change theory that the researchers were testing (e.g., increasing the salience of the perception of harm, encouraging goal setting, subjective norming) (Irvine et al., 2017). In a texting program for smoking cessation, researchers measured the frequency, length (i.e., more than 30 characters), common themes (e.g., well-being, self-efficacy, and reasons to quit smoking), and use of keywords (e.g., "stress" or "alcohol") in recipient responses throughout the intervention (Cartujano-Barrera et al., 2019). Research on broader virtual engagement contexts (e.g., online shopping, video games, etc.) has also sought to measure the emotional affect (e.g., positive versus negative) and "window" of interactions (i.e., point of engagement, period of sustained engagement, disengagement, and reengagement; O'Brien and Toms, 2008).

Even focusing on programs conducted in the education arena, definitions of text message engagement vary. Following text-based outreach to parents that promoted literacy activities for their children, York, Leob, and Doss (2018) conducted a separate survey of participants to determine whether they read and/or used the text messages, found them helpful, and shared them with other parents. In another two-way interactive text campaign designed to provide parents with personalized information about their children's school attendance, Smythe-Leistico and Page (2018) generally scanned parents' inbound text messages to a single staff member and found that most messages were questions about school schedules or requests to relay information to teachers. Castleman et al. (2017) used coarse measures of punctuation to identify the frequency of students asking questions in a financial aid filing messaging campaign. Furthermore, in a two-way texting intervention designed to reduce summer "melt," Castleman and Page (2015) looked at response rates broken out by experimental groups and sites, such as the share of students who replied to at least one text message and the share of students who replied to at least one message to request an advising meeting.

Overall, the literature to date suggests relatively little consensus around how engagement in text-based contexts is measured besides the most generic measurement of response rates. Especially as technology-enabled interventions (e.g., those delivered via SMS, email, Zoom, etc.) and the collection of these data proliferate, future research would greatly benefit from greater consistency to understand engagement as a construct across contexts. In addition, the relationships between engagement behaviors and outcomes remain understudied - i.e., whether a particular behavior is associated with or leads to a desirable outcome. This gap is surprising given participants' behavioral engagement and response are central to the success of any intervention or treatment program. This paper seeks to start bridging this gap in knowledge, specifically by identifying any correlations between specific engagement behaviors and academic outcomes like college persistence.

III. Context and Data

IIIa. Intervention Context

Nudges to the Finish Line (N2FL) was a field experiment that investigated the use of text-based nudge strategies to increase degree completion among students who had accumulated substantial credits but were at risk of withdrawal before finishing their program of study. The goal of the intervention was to increase rates of college persistence and completion. The N2FL intervention spanned several academic years, including a pilot phase during the 2016-17 academic year, followed by full-scale implementation during the 2017-2018 and 2018-2019 academic years. The Nudge[4] Solutions Lab at the University of Virginia partnered with 20 broad-access public two- and four-year institutions across Virginia, New York, Texas, Ohio, and Washington. Experimental analyses reveal the intervention did not improve student persistence or graduation rates, either in the full sample or within student subgroups (Bettinger et al., 2020).

The intervention targeted students who had **(a)** registered for classes during the first term of implementation by their institution's census date, **(b)** had a valid phone number on file, and **(c)** had previously completed at least 50% of the credits required to graduate from their program of study.

All students meeting those criteria were randomly assigned to an experimental condition, and those assigned to treatment were automatically enrolled into the interactive texting campaign. Institutional partners provided student-level administrative data (e.g., cell phone numbers and first names) necessary for delivering personalized messages.

Treated students were enrolled in their campus's texting campaign for an average of 2-3 academic semesters (excluding summer terms). They received approximately one scheduled text message per week. The messages prompted students to complete important tasks (e.g., submit the FAFSA), encouraged them to use campus resources to advance toward their degree (e.g., academic tutors, financial aid officers), and addressed feelings of stress and anxiety (e.g., financial hardships, balancing family and work). The messages leveraged behavioral insights (i.e., planning prompts, descriptive social norms, loss aversion) to increase follow-through for intended actions, and some embedded infographics (that appeared as images on students' phones) to further reinforce the call to action and increase the salience of relevant information. The research team worked with each partner campus to develop and tailor scheduled message content to their institutional context, such as inserting the specific name of a tutoring center or adjusting the tone and language for an older student population, while maintaining general consistency in content and intention across sites.

An important feature of this intervention was the ability for students to write back to the scheduled campaign messages and ask questions. The majority of the scheduled messages invited responses from students (e.g., "Registration for Spring semester starts 10/1. Want to work together to check which courses you still need to complete your degree?"), but students were able to write to their advisor at any point to initiate conversations as well.[2] This opportunity for student input and engagement is the focus of the present study. Each institution adopted a different advising model depending on the individual(s) they identified as responsible for monitoring and responding to inbound text messages from students. Campuses further adjusted the language and timing of scheduled messages to reflect the scope of those individuals' role (e.g., financial aid advisor vs. general staff assistant) and availability (e.g., to respond to students who texted in). More detailed information on these advising models can be found in Appendix VIIIa.

We note the N2FL intervention was largely restricted to text interactions between an individual student and their assigned advisor. Students may have met with an advisor or used campus resources like a tutoring center as a result of those conversations, but the core intervention consisted of two-way interactions via text message between students and a designated advisor or staff, and we only have access to message logs (and not, for example, in-person advisor visit logs) for measuring engagement in this analysis.

IIIb. Study Sample

Our analysis focuses on student-advisor interactions during the Scale Phase of N2FL that took place during the 2018-19 and 2019-20 academic years at the City University of New York (CUNY), Virginia Community College System (VCCS) and the Texas Higher Education Coordinating Board (THECB). Colleges in this phase typically implemented the intervention for 2-3

[2] While students were reading and texting from their phones, designated campus staff or advisors were reading and texting from their computers through a web-based portal called Signal Vine. Signal Vine's interface is very similar to that of an email client like Gmail or Microsoft Outlook. A couple benefits of the Signal Vine system in this project included filtering which inbound messages from students were unread and required follow-up, and scheduling messages for future delivery (e.g., if the advisor or staff wanted to schedule a reminder message about an upcoming scholarship deadline that was still a couple weeks away).

academic terms, not including summer terms (e.g., sending messages during spring 2018, fall 2018, and spring 2019 semesters for a three-term schedule) across multiple cohorts (e.g., at one college, the first cohort's messages spanned spring 2018 through spring 2019 semesters, while the second cohort's messages spanned fall 2018 through spring 2019).

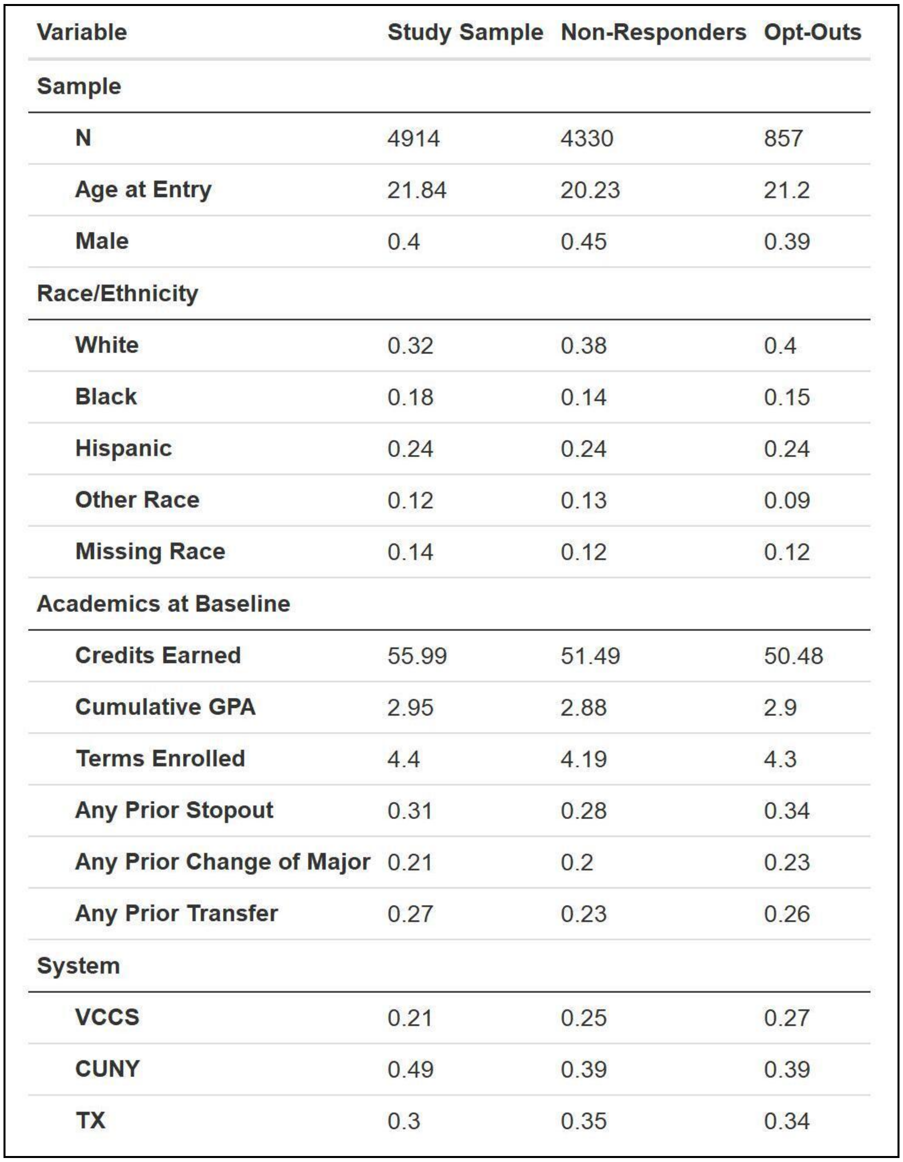
For the purposes of this study on patterns of student engagement, we focus explicitly on the texts sent by students. The full message log transcripts include roughly 327,000 texts sent or received throughout the intervention, and about 34,000 (~10%) were sent by students.[3] We moreover focus on students who responded to at least one scheduled message and did not request to opt-out of texting at any point during the intervention. This is to focus on patterns of engagement among those students who *actually* engaged in the study.[4] This results in a total of 4,914 students in our sample and 33,177 student texts.

Table 1 reports the demographic and academic baseline characteristics of each respondent group: our analytic sample, students who never responded to any scheduled message, and students who requested to opt-out of texts. First, we note that about half of treated students either opted-out (8.5%) or never responded (42.9%), indicating that the actual take-up of any text-based advising was fairly low (48.6%). In general, we see that the three samples are largely similar to one another with only minor differences. Students in the analytic sample were only slightly older than the other groups. About 40% of the analytic sample of responders were male relative to 45% of non-responders and 39% of opt-outs. Our study sample was also more likely to be Black (18%) than non-responders (14%) or opt-out students (15%). Analytic sample students seemed to be slightly higher in academic performance than other groups, with more credits earned at baseline, slightly higher cumulative GPA, and more terms enrolled. Study students were more likely to have had a stopout (unenrolled from the college without graduating) before the intervention than non-responders, but less likely to have had a stopout relative to the opt-outs. Sample students were otherwise about as likely to change their major or transfer prior to the intervention start date as the other groups. Finally, we see that VCCS and TX students were slightly more likely to be a non-responder or opt-out than included in the sample, whereas CUNY students were substantially more likely to be included in the sample.

Table 1. Demographic and Academic Baseline Characteristics by Respondent Group

[3] If treated students texted their advisors after the full set of scheduled messages were sent, they received an automated response notifying them that the campaign had concluded. To keep the timeframe of interest in the present study consistent across cohorts (and to avoid including any texts sent during the COVID-19 pandemic), we exclude any texts sent to or from students more than 14 days after the last scheduled message in the intervention was delivered. This excludes 104 texts by students and 60 texts by advisors.

[4] Subsequent analyses will examine the extent to which design features of the intervention affected likelihood of engagement as well as patterns of engagement.



Variable	Study Sample	Non-Responders	Opt-Outs
Sample			
N	4914	4330	857
Age at Entry	21.84	20.23	21.2
Male	0.4	0.45	0.39
Race/Ethnicity			
White	0.32	0.38	0.4
Black	0.18	0.14	0.15
Hispanic	0.24	0.24	0.24
Other Race	0.12	0.13	0.09
Missing Race	0.14	0.12	0.12
Academics at Baseline			
Credits Earned	55.99	51.49	50.48
Cumulative GPA	2.95	2.88	2.9
Terms Enrolled	4.4	4.19	4.3
Any Prior Stopout	0.31	0.28	0.34
Any Prior Change of Major	0.21	0.2	0.23
Any Prior Transfer	0.27	0.23	0.26
System			
VCCS	0.21	0.25	0.27
CUNY	0.49	0.39	0.39
TX	0.3	0.35	0.34

IV. Methodology

IVa. Analytic Framework

The broader N2FL evaluation examined the effect of enrolling students in the text-based advising, and we can think of the overall treatment experience as the bundle of scheduled messages (e.g., helpful reminders about upcoming deadlines) and the availability of text-based advising (i.e., personalized support that students could access via text message). This latter aspect of two-way interaction was one of the most distinctive features of the N2FL campaign, setting it apart from one-way informational texting campaigns.

We might expect the availability of this text-based advising to impact student outcomes through two separate, but related, mechanisms. The first mechanism is what we herein refer to as "active engagement," where students benefit from the personalized support and information they receive specifically by way of actively engaging with the text-based advising. We then consider variation in students' active engagement a meaningful source of *treatment intensity* in the broader intervention. That is, students who engaged heavily with advisors via texting would have received greater levels of this text-based advising "treatment" in the intervention than students who did not, and variance in these greater levels of engagement may then relate to variance in student outcomes.

Text-based advising may also affect student outcomes through a second mechanism we herein refer to as "passive engagement," in which students benefit specifically from the mere knowledge that they are being supported in the abstract. In other words, just knowing that an advisor is at their fingertips may confer some psychological benefits to students, such as a heightened sense of belonging in the college community that helps them perform better in their classes and engage in adaptive behaviors (Gopalan & Brady, 2019).

In contrast with the N2FL evaluation that examines the causal effect of the combined bundle of scheduled messaging, passive engagement, and active engagement, our analysis focuses on the characteristics, predictors, and correlates specifically of *active* engagement in the N2FL intervention to the extent possible.

Directly assessing the causal impact of active engagement on student outcomes in this context is complicated for two reasons. First, active engagement is not randomly assigned across students. Variance in active engagement rates is primarily driven by students' decisions to engage with the advising platform (and, to a lesser extent, the responsiveness of individual advisors), and active engagement rates may then be endogenous to observable and unobservable student characteristics. We thus cannot estimate a causal relationship between engagement and student outcomes using these observational data without our results being contaminated by omitted variables bias from unobservable student characteristics (e.g., students who are more conscientious are both more likely to respond to any texts they receive *and* perform well in their classes). Moreover, because we cannot know the engagement levels of students in the control group *had they been sent texts*, we cannot leverage the original intervention's RCT design to estimate the impact of varying levels of engagement, either.

Second is the measurement issue. Engagement (as we described earlier in Section II) is a multi-dimensional and complex construct, and it is not clear which measurable proxies best stand-in for active engagement of text-based advising itself. For example, the N2FL texting data did not capture whether a scheduled message was actually read by students, and so we cannot know how many messages a student read. While we can look at the number of responses a student sent, a single-word text response from a student (e.g., "Yeah") would still be conceptually different from a more involved, inquisitive response (e.g., "Yeah, and I was hoping you could tell me more about…") when thinking about engagement and treatment intensity.

These considerations set the stage for our study in two parts. We begin by exploring a variety of possible engagement measures to document variation across students and any correlation with one another, as guided by the literature when possible. Our driving motivation is to generate novel insight into student engagement patterns that are broadly applicable across intervention contexts; to this end, we choose to focus on measures that are observable and conceptually relevant to the construct of engagement (i.e., behavioral dimensions like length of engagement periods), that can be easily communicated to and applied by other researchers, that document meaningful levels of variation across students, and that are generally uncorrelated with other measures. We interpret those measures with large levels of variation as those more likely to be malleable, though this dynamic will need to be explored in more detail in future work. We then examine whether any of these engagement measures are also correlated (descriptively) with the outcomes of interest. Ultimately, we are attempting to identify promising proxies for student engagement in these text-based advising campaigns that are associated with better student outcomes, thus generating testable hypotheses about which engagement measures are most malleable and impactful. These hypotheses would then open pathways for future work to explore the manipulation of such engagement measures in experimental contexts as we seek to improve the efficacy of text-based advising programs.

IVb. Defining Engagement

We break our student-level engagement measures into four main categories as guided by the literature and the data we have at our disposal. Rather than attempting to derive a single measure of engagement, we attempt instead to create an array of measures that captures the various nuances of engagement across several dimensions.

In the first category, we are interested in examining the **Frequency and Intensity of Student Replies**. This category most closely mirrors how response rates have been measured in related literature. Because some sites for the N2FL intervention chose to send their scheduled messages at differing intervals and frequencies (e.g., due to different dates for financial aid filing, course registration, etc.), we examine the *percent of scheduled messages that a student responded to at least once* This can be thought of as the relative frequency with which a student had *any* apparent engagement in the scheduled messages they received, meaning it does not consider circumstances when students have sustained engagement after a given scheduled message. To address that shortcoming of the measure, we also examine a student's *average replies per scheduled message* (inclusive of any messages students send in response to their advisor afterward), to assess how frequently a student engaged in a more sustained way with the scheduled messages. To examine engagement intensity, we also look at a student's *average reply length in words* under the assumption that longer messages are indicative of a more intense level of engagement with their advisor.[5] Similarly, we examine a student's *proportion of substantive replies*, or the share of replies at least 5 words in length.[6] This is to distinguish quick, gestural answers (e.g., "Yes", or "No") from more deliberate responses.

We also are interested in examining **Engagement Duration** as a separate category. For example, consider two students who responded to exactly four scheduled messages: Student A responded to the first four scheduled messages of the campaign and then disengaged completely, and Student B responded to four scheduled messages of the campaign throughout a period of three academic semesters. We thus calculate the *percent of scheduled messages sent before a student engaged*(i.e., sent their first reply) and the *percent of scheduledmessages sent before a student disengaged* (i.e., sent their last reply). We then also calculate the share of scheduled messages sent within this window, which we interpret as the *percent of scheduled messages a studentwas engaged.*

Another category of interest for student engagement behaviors that is measurable across most texting intervention contexts is their **Response Speed**. Perhaps students who are quick responders paid closer attention to the intervention than those who spent several hours or days before replying. Conversely, it is possible that students who responded later spent more time thinking about and internalizing the scheduled messages before responding. To measure this dynamic empirically, we first calculate a student's *average response time to scheduled messages* in hours (conditional on responding). We also calculate a student's *average response time to any advisor-generated messages* in hours (conditional on responding) to also capture a student's reply speed to

[5] We also examine reply length in *characters* as an alternative specification and find no substantive difference to our results. We focus on words in this manuscript for its interpretability and concision.

[6] This threshold for 5 words is based largely on our informal examination of short texts in the data, where we find that responses below 5 words tend to be gestural or confirmatory in nature and without other substance.

non-scheduled advisor-generated messages (e.g., messages their advisor directly wrote to them as part of the conversation).

Finally, we are not just interested in the patterns of when and how often a student responds, but also *what they discuss* when they do. To that end, we employ a range of text analysis techniques to better understand and measure **Response Content**. The first technique is relatively straightforward where we scan each student's text for the prevalence of a given language category. For example, this approach would scan messages for "help-asking" language – messages with question marks, as well as messages including phrases like "how do I…," how do you…," "can you…," "I need…," and so on. We can then calculate the *proportion of studentmessages asking for help* as a specific type of engagement relevant to the text-based advising context of the N2FL intervention.[7]

We go on to create two more complex text content measures using NLP techniques known as sentiment analysis and structural topic modeling. For concision, we provide only a brief and intuitive explanation of these two methods in the next sections of the main paper; greater detail on the techniques themselves, their implicit assumptions, our text cleaning decisions, the robustness checks we ran, and the validity exercises we deployed, can be found in Appendices VIIIb-VIIIf.

IVc. Analyzing Message Content with Sentiment Analysis

Sentiment Analysis is a common NLP task in which analysts use an algorithm to "read" a given text string and rate the extent to which the string contains/expresses a positive, negative, or neutral/factual sentiment (Pang & Lee, 2008). One can think of this process as generating output similar to hand-coded qualitative analysis, but in an automated and highly scalable way that facilitates quantitative analysis. Though matching human judgment *perfectly* remains out of reach for the current state-of-the-art, modern algorithms are nonetheless exceptionally nuanced, flexible, and accurate[8] at this task (Vaswani et al., 2017). Using vast volumes of text data to first "learn" a general understanding of language syntax and word relationships, modern sentiment analysis algorithms are now able to account for the complexities of word context (e.g. that "I wish I were happy" actually indicates sadness), multiple word meanings (e.g. that "bank" has two separate meanings in "The river bank was wet" and "I went to the bank this morning"), and informalities (e.g. "that was sick, dude;" Ambartsoumian & Popowich, 2018) far better than early sentiment analysis approaches (e.g., dictionary-based methods).

We operationalize the definition of sentiment for the present study as the *perceived positivity of emotions and ideas present in a given text.* This definition then is a conglomeration of the speaker's stated emotions ("I feel sad" v. "I am excited"), communicated intention ("I hope you die" v. "I wish you the best!"), and, at least to some extent, topical content ("financial hardship" v. "vacation time"). Note that this definition is complicated for longer strings of text, in which multiple emotions/implications may be present and the overall sentiment becomes more ambiguous.[9] We

[7] This approach also enables us to identify students who opted out from the intervention, searching for phrases such as "stop messaging" or singular responses consisting only of "sotp."

[8] Importantly, it is often the case that a given string of text has no one "right" answer for its sentiment, and so expecting an algorithm to perfectly match human judgment may be an impossible bar to set to begin with. In our own validation exercises, for example, our team of five human coders often disagreed about a given text's sentiment due to individual interpretations of subtext and implications. We ultimately find that the inter-rater reliability of a team of human coders is not significantly different from the inter-rater reliability of a team of human coders*plus the algorithm*, indicating that it does not seem to disagree with a human's judgment about a given text any more than humans disagree with one another.

[9] While we would like to lean on a more standardized definition, we were unable to find a detailed and widely-accepted definition for sentiment in transformer-based models in the literature. Interestingly, sentiment as a construct across

argue this definition remains appropriate for our context because of the nature of the N2FL text data: texts were generally only one or two sentences long (making the ambiguity of sentiment in long text strings less problematic), students are sending these texts "as-is" (i.e., they are not transcribed spoken words with greater context than what we can observe in the text data), and we need not perfectly describe the student's *intended* sentiment for this to nonetheless be a useful typology for classifying distinct modes of engagement. For demonstration purposes, we provide a list of real N2FL texts from students alongside their algorithmically-generated sentiment score in Table 2 below.

Table 2. Sample N2FL Messages and Assigned Sentiment Scores

Text Message (sic)	Sentiment Score
Terrible I have to find a class or two to sign up for. I'm so behind.	Very Negative (-2)
No. Everything has been piling up at school and it's kind of been too stressful to decide what to get done.	Very Negative (-2)
I was trying to drop a class and it doesnt allow me	Negative (-1)
Hi! I applied for graduation and got an update but I do not understand it because it doesn't match the update list on the [institution] graduation page	Negative (-1)
I am in school Tuesdays and Thursdays	Neutral (0)
When does summer classes start?	Neutral (0)
Just tried again and it let me register haha, thank you for your help	Positive (1)
It's fine. Thank you very much for the link. If I have any other questions in the future, can I text this number?	Positive (1)
Ok thanks so much! I finished this semester strong! I got a 100.5% on my Anatomy Final Exam! That grade replaced my lowest test grade of an 85%	Very Positive (2)
It was very helpful thank you so much!	Very Positive (2)

Note: Texts shown here were specifically selected from the set of N2FL texts where the human coders and the algorithm output were in agreement to clearly illustrate what differing levels of sentiment can look like. These examples should not be interpreted as a general demonstration of algorithm accuracy.

We ultimately measure the *average positivity of emotionsand ideas present in a student's replies* (average "sentiment") on a scale from very negative (-2) to very positive (2), where sentiment with a score of 0 can be thought of as more factual in nature (e.g., "I enrolled in my courses" rather than "I was so relieved to enroll in my courses"). In other words, what is the general tone across a student's responses?

modern data science (i.e. neural network-based models rather than dictionary models) is almost entirely dependent on the SST's definition due to the strong incentive for data scientists to optimize their algorithm's SST performance for benchmarking purposes. That said, the SST intentionally encouraged their human coders to view sentiment as a flexible and subjective notion, making a formal definition elusive.

IVd. Analyzing Message Content with Topic Modeling

Topic Modeling is another common task in NLP in which analysts use an algorithm to "learn" what discrete topics of discussion exist across a series of text documents (in our case, texting conversations) and then measure how prevalent each topic is within each of the provided documents. In brief, the algorithm does this by examining how words are used in conjunction with one another across documents, under the assumption that words about the same general topic of conversation will often appear together in the same documents (Blei, 2003). For example, "financial," "aid," "deadline," and "FAFSA" might often appear in the same documents, thus indicating to the algorithm that they are used to discuss the same topic of conversation. By then constructing several sets of words that often appear together in this way, the algorithm will have identified the word groups that it thinks represent each distinct topic of conversation within the text data; analysts then interpret these word groupings for meaning (such as, "FAFSA filing") and, in our case, make "supertopics" that combine multiple word groups together under a single broader category of conversation (such as, "financial aid" that encompasses topics about FAFSA filing, tuition payments, etc.).

Ultimately, we are interested in whether there exists variation in the prevalence of these supertopics across students' replies. Such variation would reflect substantively different engagement behaviors, and thus different patterns of *how* students navigate their responses to the advising intervention as a result. Once the supertopics are identified, the algorithm can determine the prevalence of each supertopic across a given student's texts based on the combination of keywords they used. For example, the algorithm can tell us how many keywords in a given student's responses are spent discussing the "financial aid" as a supertopic, versus "course planning." We thus construct measures to describe the *percent of student replies*about each of the following topics: course planning (e.g., course registration, registration deadlines, etc.), financial aid (e.g., applying for financial aid, paying tuition, etc.), academic planning (e.g., graduation deadlines, transfer requirements, career planning, etc.), general academic support (e.g., tutoring services, study skills, etc.), and meeting logistics (e.g., scheduling an in-person advising meeting, getting the right contact email, etc.). In other words, what are the more prevalent topics of conversation in a student's responses? A partial list of the most influential keywords that fall under each supertopic is displayed in Table 3 below. A complete list can be found in Appendix Table F2.

Table 3. Supertopic Groupings and Sample Subtopics and Words

Academic Planning	math, science, requirement, biology, art, spanish, registrar, language, college
	credit, graduate, course, major, requirement, internship, psychology, minor
	graduate, congratulations, graduating, applied, feel, free, ready, december
	degree, transfer, major, change, associates, plan, audit, transcript, bachelors
Academic Support	hope, information, tokenurl, hey, center, tutoring, located, helpful, office, visit
	question, hey, info, yeah, answer, reaching, nice, assist, specific, study
	im, semester, grade, luck, checking, enrolled, final, exam, planning, lol
	campus, service, counselor, job, support, mind, ahead, provide, care, set
Meeting Logistics	appointment, time, tomorrow, wednesday, thursday, tuesday, monday, meet
	message, office, time, answer, frame, time frame, message time, answer message
	tokenphonenumber, call, phone, person, walk, call tokenphonenumber, monday
	advisor, contact, academic, tokenname, meet, advising, academic advisor, track
	appointment, schedule, schedule appointment, set, advising, advisor, tokenurl
Course Planning	spring, registration, date, winter, spring semester, november, enrollment, session, register
	student, id, drop, time, gpa, student email, check, access, withdraw, student id
	summer, fall, course, taking, summer class, online, fall semester, summer course
	professor, department, told, writing, speak, permission, request, alright, issue
Financial Aid	tokensis, hold, account, plan, payment, pay, bursar, log, check, tokenurl
	financial, aid, financial aid, fafsa, office, aid office, scholarship, loan, tuition, pay

Note: We display the top ~10 words within each sub-topic in terms of its*probability* metric (how much an appearance of that word contributes to the detection of that topic). We display only the first four subtopics under each supertopic for concision – a full list of the subtopics can be found in Appendix Table F2. “tokenname, “tokenphonenumber,” “tokensystemname,” and so on, were placeholders used for scrubbed PII words.

IVe. Regression Analysis

Besides examining how students vary in terms of their behavior across the engagement measures, we are especially interested in the extent to which these engagement measures are related to the outcomes of interest for the intervention: re-enrollment term-to-term (binary), credits earned (number of credits), term GPA (raw GPA units), and degree receipt (binary). In other words, do we see any relationship between how students engage and their ensuing academic performance? Because this is a descriptive analysis, we again cannot take any of these relationships as causal; instead, we think of this as an exploratory analysis meant to generate testable hypotheses for future research (e.g., experiments) around what engagement measures might be both malleable and also impactful on outcomes. For example, if it is the case that students who ask for help more frequently via text tend to perform significantly better in terms of desired student outcomes, researchers might explicitly explore this relationship further in future work by designing a text-based advising intervention with prompts greater or fewer student questions across treatment wings to see if such intervention features enhance the effectiveness of said intervention.

We examine the relationship between each engagement measure and each outcome of interest (in the first, second, and third term immediately following the start of the N2FL intervention) in the context of a regression analysis where we also control for salient student demographics and baseline academic characteristics. For controls, we include all of the variables explored in Table 1, as well as the randomization block of students during the initial study randomization process.[10] More formally, we iteratively estimate the following equation:

$$(1) \quad Y_i = \lambda_t + X_i + A_i + \beta_1 Engagement_i + \varepsilon_i$$

where Y_i represents any one of the student outcomes of interest, λ_t represents the vector of fixed effects for student randomization blocks, X_i represents the vector of student demographic characteristics, A_i represents the vector of student baseline academic characteristics, $Engagement_i$ is any one of the engagement measures, and ε_i represents the idiosyncratic error term. The coefficient of interest will thus be beta 1, revealing the controlled relationship between each engagement measure and each outcome. Finally, we cluster our standard errors at that randomization block level.

V. Results

Va. Variation in Engagement Measures Across Students

We first look at the distribution of each measure at the student-level to understand how they vary across students and potentially uncover salient patterns of engagement in the intervention. In all following plots, the X-axis charts out the range of the values for a given engagement measure, the Y-axis shows the density of students at each value along the X-axis, the dotted line shows the mean value of the engagement measure at the student-level (also reported in the subtitle), the solid line shows the median value of the engagement measure at the student-level (also reported in the subtitle), and the number of students displayed in each plot (i.e., number of non-missing values) is indicated in the subtitle of each plot.

Figure 1 shows the distribution for each of the **Frequency and Intensity of Student Replies** measures. In the top-left panel, we see the vast majority of students who responded to *any* message still responded to fewer than 25% of the scheduled messages they receive, with a long tail extending beyond that. In the top-right panel, we see that most students only respond with a single reply after a given scheduled message on average, indicating that they very rarely engage in actual back-and-forth texting with their advisors when they do respond. The long tail here also indicates that a small handful of students regularly engaged in lengthier conversations. These metrics point to the reality that, even in well-designed interventions seeking to elicit two-way student engagement, genuine student engagement may be less common than we might otherwise expect. In the bottom-left panel, we note that the student-level median for average reply length is 10 words – the equivalent of a short sentence, which makes sense given the medium of texting. Interestingly, the

[10] Note that the randomization blocks were separated by school system, meaning school system is completely collinear with randomization block across all students and is thus unnecessary to include separately in this regression.

highest density area of students always responded with a substantive reply (>=5 words), as shown in the bottom-right panel, while a far smaller proportion never did. That the student-level median for the proportion of substantive replies is 0.72 also indicates that insubstantial replies (e.g., "yeah," "okay," "no," "thanks", etc.) were less common than we might have initially anticipated a priori for a texting intervention.

Figure 1. Distribution of Engagement Measures: Frequency and Intensity of Student Replies

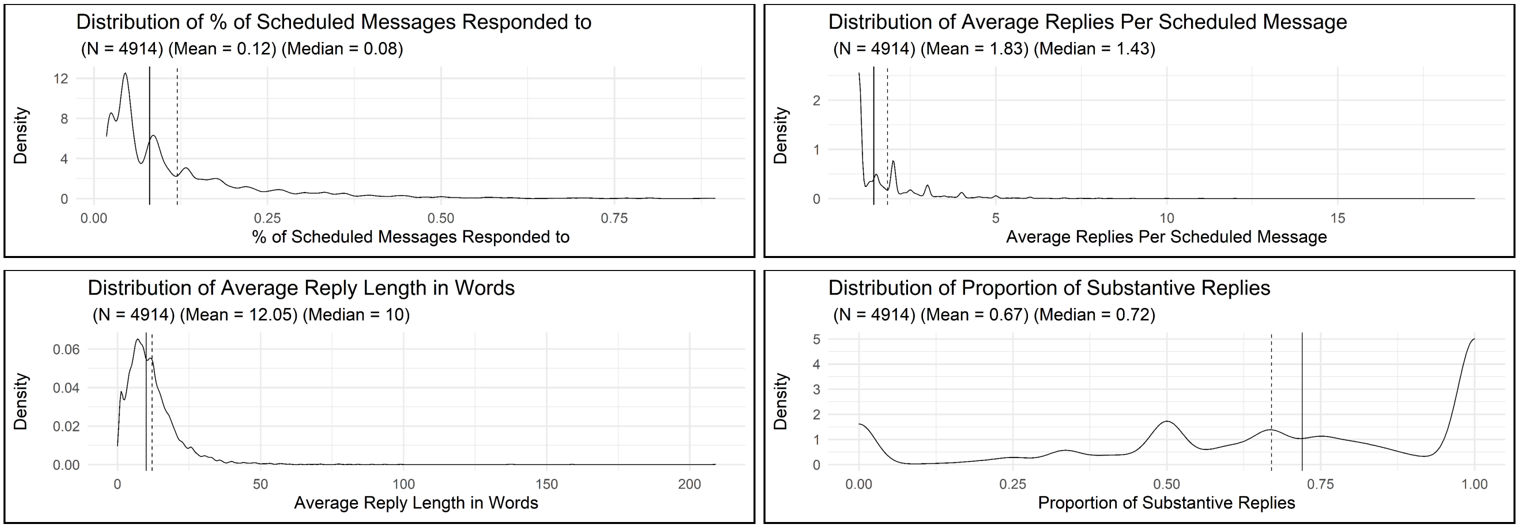


Figure 2a shows the distribution for each of the **Engagement Duration** measures. In the top-left plot we show the percent of messages sent prior to students' first response - the mass near zero indicates that most students engaged for the first time very early on in their scheduled messages, with a student-level median of 0.17. That said, a long tail here also indicates that a meaningful share of students were nonetheless engaging for the first time all along the sequence of scheduled messages. The next plot reveals also that the median student disengaged (i.e., sent their last reply) about two-thirds of the way through the intervention, though many didn't disengage until the very end given the mass around 1. Finally, in the bottom plot we show the distribution of the percent of messages with which students engaged. The mass of points near zero in the bottom plot reveals that, despite the distributions showing many early engagements and many late disengagements, relatively few students were engaged for the majority of the intervention. The mass near zero indicates that a large share of students were only ever engaged for a brief period of the intervention.

Figure 2a. Distribution of Engagement Measures: Engagement Duration

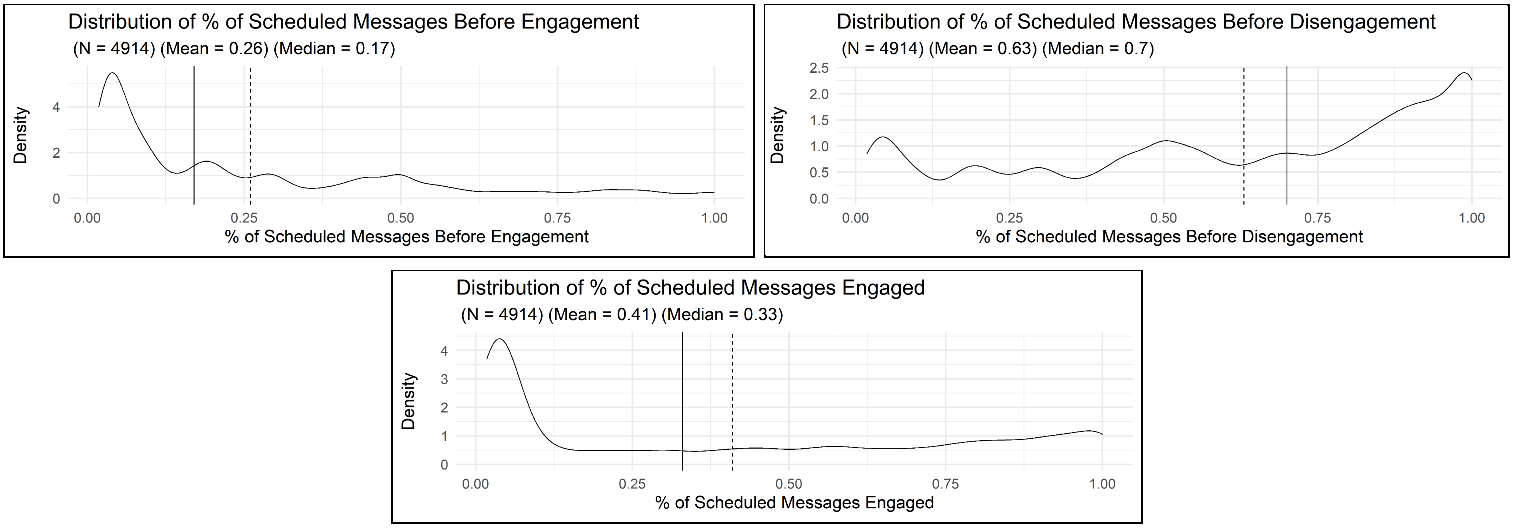


Figure 2b is another way of visualizing the same engagement duration data to better differentiate individual students' behaviors and explain these seemingly contradictory results. Along the X-axis is the percent of scheduled messages before students sent their first message, while along the Y-axis is the percent of scheduled messages before students sent their last message. We can then, for an individual student represented as a single point, see when they engaged relative to when they disengaged. As an example, students in the top-left of the plot engaged immediately at the start of the intervention (the percent of scheduled messages that passed before they engaged was nearly zero) and disengaged at the very end of the intervention (the percent of scheduled messages that passed before they disengaged was nearly one), indicating that they were engaged throughout the entirety of the intervention (the percent of scheduled messages they were engaged for was 1). Students along the 45-degree line are students who engaged and disengaged at the same time, and thus must have only ever sent one reply.

Overall, we see that many students could only nominally be considered actively engaged at all under this definition given the mass of points along the 45-degree line. That said, a fair cluster of students in the top-left and along the left side of the plot had immediately engaged and stayed engaged for a large proportion of the intervention, reflecting the long tail of students in the prior plot for the percent of scheduled messages they were engaged (bottom plot of Figure 2a). Lastly, students were most likely to engage early on in the intervention if at all, given the decreasing number of points as we move along the X-axis, reflecting the large mass of points near zero in the distribution of percent of scheduled messages before students engaged (top-left plot of Figure 2a).

Figure 2b. Scatterplot of Engagement Duration Patterns

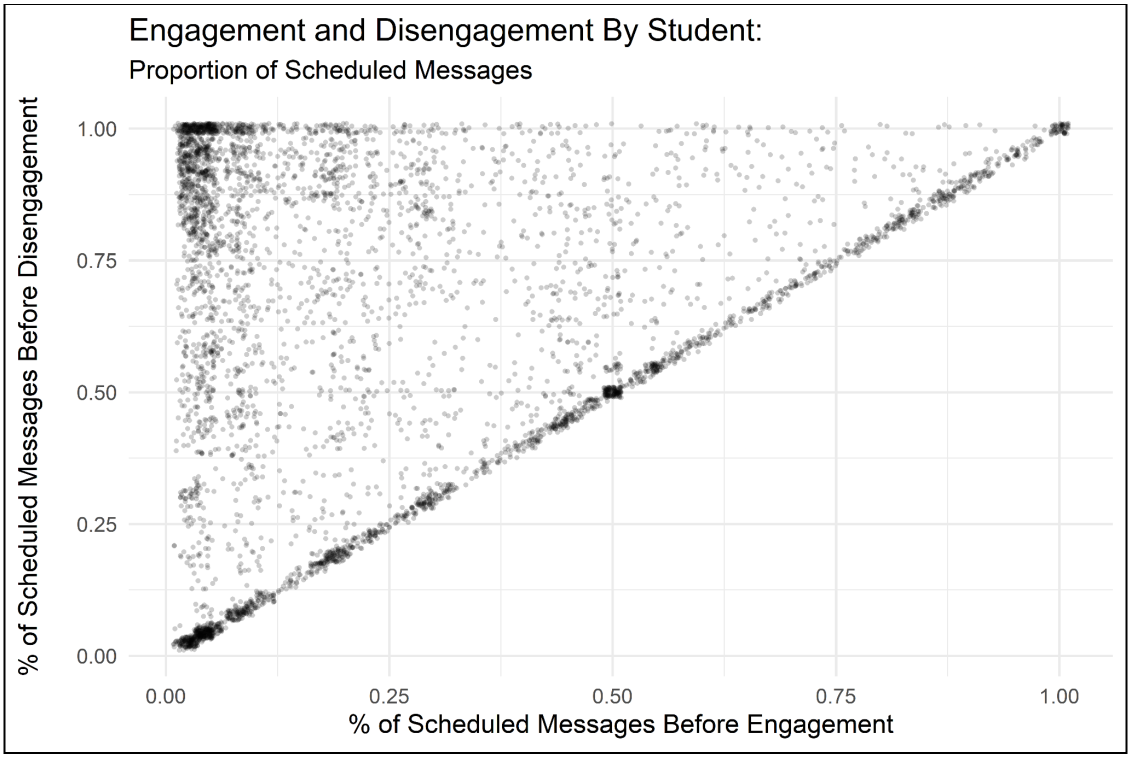
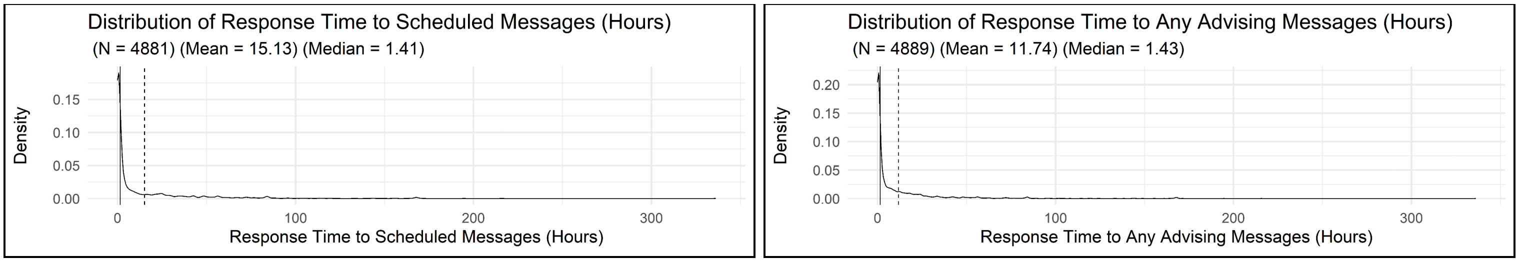


Figure 3 shows the distribution for each of the **Response Speed** measures.[11] The left plot reveals that the median student responded within 1.41 hours of scheduled messages, though there exists an exceptionally long tail where students replied days, or even weeks, after most scheduled messages; we might interpret this to mean that, even if students did not reply promptly to scheduled messages, they were aware of the availability of text-based advising and turned to this mode of communication days and weeks further out. We do not see that this distribution pattern changes meaningfully when more broadly considering students' responses to *any* advising texts (e.g., ad hoc messages that advisors wrote in response to students' initial replies), though the average response time is slightly reduced from 15.13 hours to 11.74 hours.

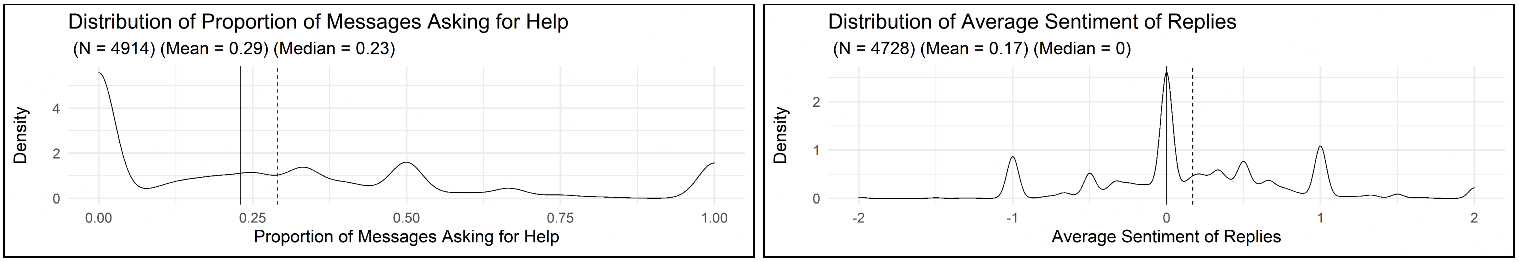
Figure 3. Distribution of Engagement Measures: Response Speed



[11] Note that while response times are reported with hours as the unit, all calculations are accurate to the second (i.e., measurements were not rounded to whole hours).

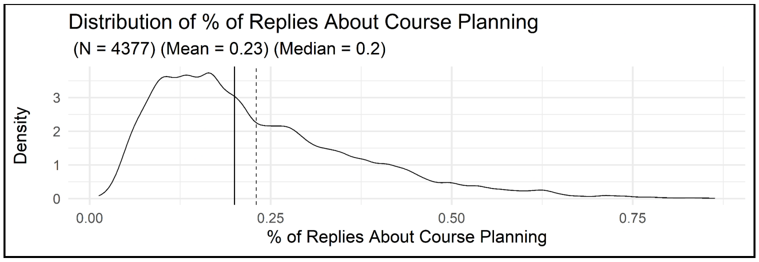
Finally, Figures 4a and 4b show the distribution for each of the **Response Content** measures, beginning with the help-asking and sentiment measures in Figure 4a. The left panel shows that there was relatively wide variation in the proportion of messages each student sent asking for help, though a large proportion of students never asked for help given the mass near zero. Paired with the fact that most student replies were about a sentence in length (bottom-left plot of Figure 1), students seemed to more often be answering questions in the scheduled messages with declarative statements than with explicit requests for additional help. We also see this dynamic reflected in the sentiment analysis results, with most students showing an average sentiment of replies at 0, suggesting the prevalence of factual statements was greater than emotionally-charged student messages (e.g., those remarking on difficulties, frustration, excitement, etc.). That said, we still see meaningful variation around 0 in both directions, so students were still sending texts with more positive and more negative sentiment.

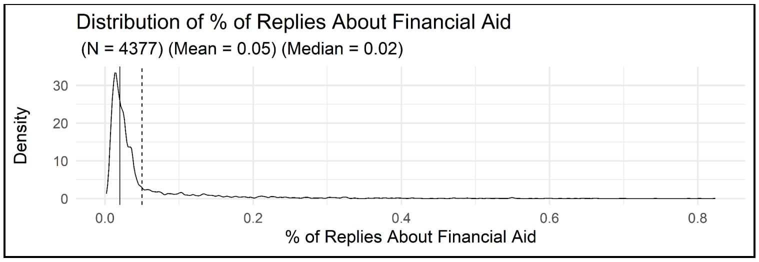
Figure 4a. Distribution of Engagement Measures: Help-Asking and Sentiment Response Content

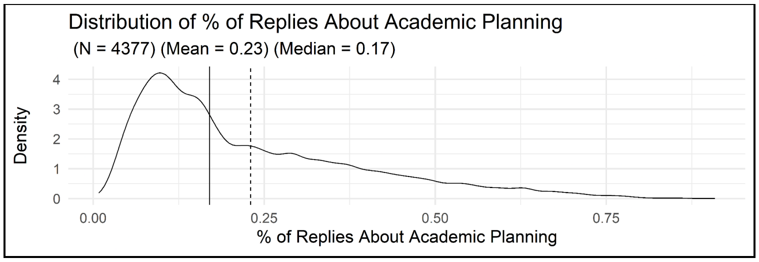


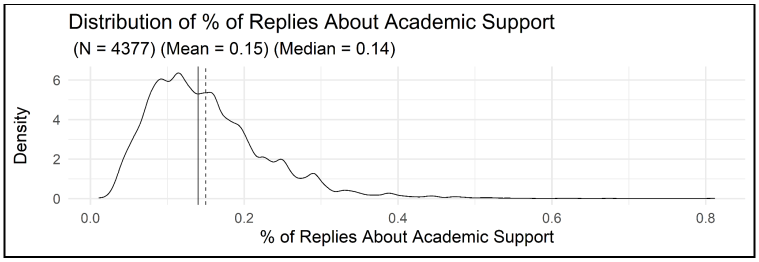
Turning now to the topical content measures in Figure 4b, we see generally wide distributions of topical content prevalence for every supertopic *except* financial aid in the top-right plot. That is, the median student spent between 13% and 20% of their replies focused on each topic of course planning, academic planning, academic support, and meeting logistics, but only 2% of their replies focused on financial aid. The wide distributions for all but financial aid indicate that students seemed to generally vary quite a bit in terms of how much they discussed each topic. This variation perhaps reflects one of the strengths of an advising intervention in that it is responsive to individual students' needs and interests. That relatively few of the replies focused on financial aid is puzzling but could be the result of a few likely dynamics: **(a)** the deadline-dependent nature of FAFSA filing and tuition payments means they might be relevant only during specific timepoints of the year, whereas the other topics could more naturally come up throughout the entirety of the intervention; **(b)** the scheduled messages on the topic instigated responses that were more confirmatory in nature (e.g., "Have you filed your FAFSA yet?") and thus didn't require students to respond using financial aid phrases; **(c)** students may have been less comfortable raising financial aid questions or issues via text message; and/or **(d)** students targeted by the intervention are near-completion, and so may already be familiar and comfortable with financial aid filing processes by this point in their college trajectories. In any case, the relatively low level of student replies about financial aid is surprising given the substantial proportion of lower-income demographics at the broad-access institutions represented in the sample.

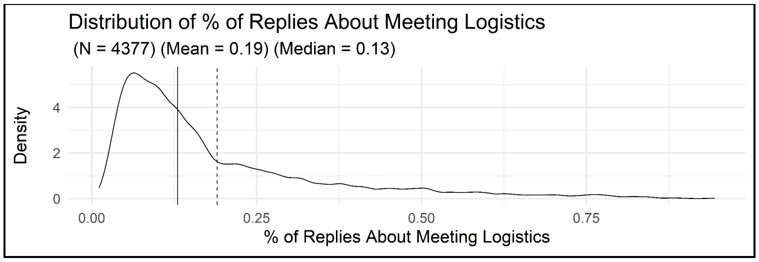
Figure 4b. Distribution of Engagement Measures: Topic Modeling Response Content











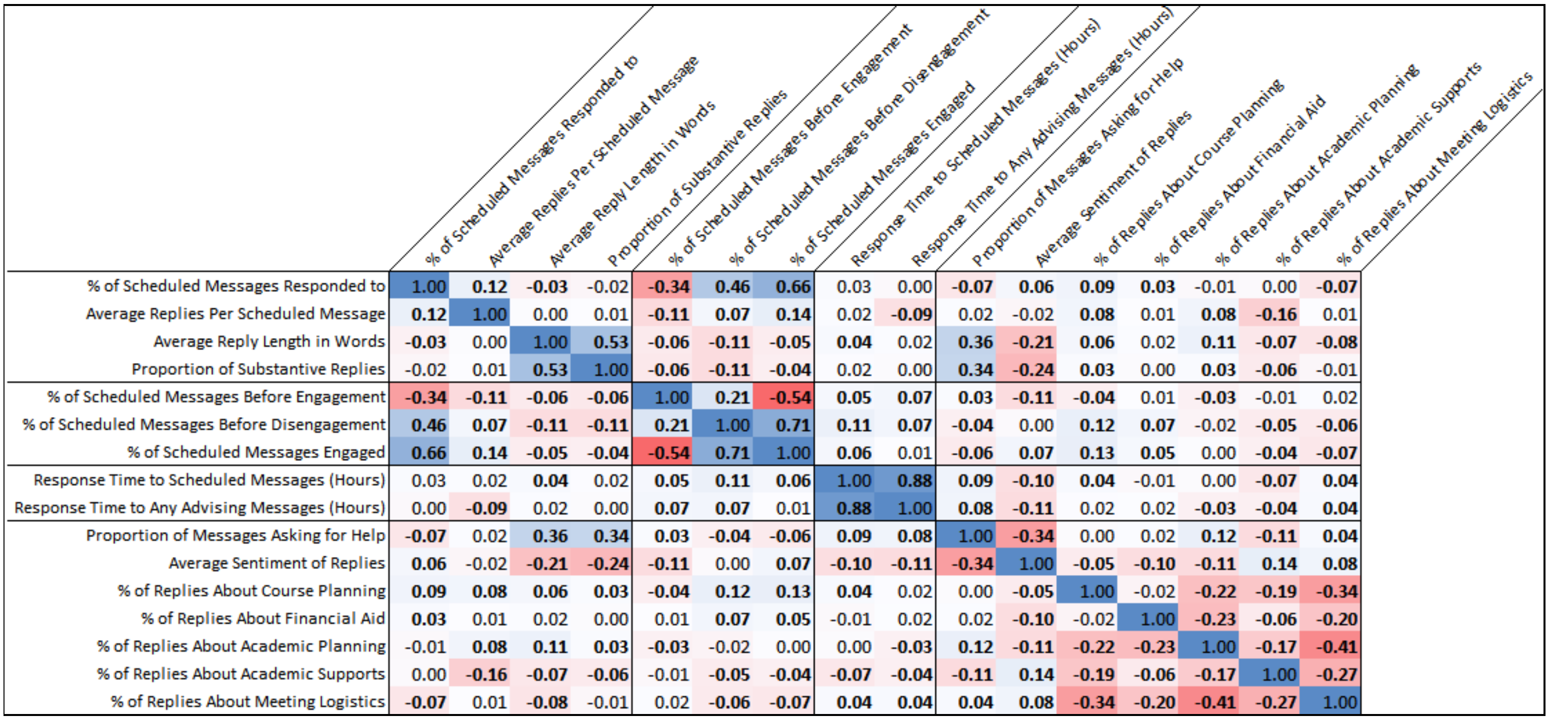
To summarize at a high level the aforementioned results, we generally see the least variation across responsive students in terms of their response times and their average responses per scheduled message. We see the greatest variation in terms of the topical content of their replies, and still meaningful variation in terms of the proportion of their messages asking for help, the sentiment of their replies, when they engaged and disengaged, and the length of their replies in words. While each measure helps reveal useful insights about patterns of student behavior (e.g., that most students respond almost immediately after they receive scheduled messages, if at all), these measures with greater levels of variation are most likely to help us distinguish students' engagement patterns from one another (e.g., versus measures with low variation such as response speed).

Vb. Correlations Between Engagement Measures

While we can learn much about how students engaged in the intervention by examining each of the measures individually, we are also interested in the extent to which these measures correlate with one another. Measures that have high levels of absolute correlation with one another can reveal "bundles" of common engagement patterns in the context of text-based advising interventions, while measures that have low levels of absolute correlation might best be interpreted as measures that capture distinct information from one another. The former might be especially useful to gain deeper insight into how students navigate text-based advising interventions in general (e.g., to inform future program design), while the latter might be especially useful as we seek to create a parsimonious set of engagement measures we could commonly track and/or encourage across text-based advising interventions of this kind.

Table 4 is a correlation matrix that shows the correlation coefficient between each of our engagement measures against one another. Cells are shaded according to their correlation, with red shading indicating stronger negative correlations and blue shading indicating stronger positive correlations. Any coefficients presented in bold are statistically significant at the p<0.05 level. Borders are drawn around each of the four groups of measures (Frequency/Intensity, Duration, Response Time, Response Content) for visual clarity.

Table 4. Correlations Between Engagement Measures



	% of Scheduled Messages Responded to	Average Replies Per Scheduled Message	Average Reply Length in Words	Proportion of Substantive Replies	% of Scheduled Messages Before Engagement	% of Scheduled Messages Before Disengagement	% of Scheduled Messages Engaged	Response Time to Scheduled Messages (Hours)	Response Time to Any Advising Messages (Hours)	Proportion of Messages Asking for Help	Average Sentiment of Replies	% of Replies About Course Planning	% of Replies About Financial Aid	% of Replies About Academic Planning	% of Replies About Academic Supports	% of Replies About Meeting Logistics
% of Scheduled Messages Responded to	1.00	**0.12**	**-0.03**	-0.02	**-0.34**	**0.46**	**0.66**	0.03	0.00	**-0.07**	**0.06**	**0.09**	**0.03**	-0.01	0.00	**-0.07**
Average Replies Per Scheduled Message	**0.12**	1.00	0.00	0.01	**-0.11**	**0.07**	**0.14**	0.02	**-0.09**	0.02	-0.02	**0.08**	0.01	**0.08**	**-0.16**	0.01
Average Reply Length in Words	**-0.03**	0.00	1.00	**0.53**	**-0.06**	**-0.11**	**-0.05**	**0.04**	0.02	**0.36**	**-0.21**	**0.06**	0.02	**0.11**	**-0.07**	**-0.08**
Proportion of Substantive Replies	-0.02	0.01	**0.53**	1.00	**-0.06**	**-0.11**	**-0.04**	0.02	0.00	**0.34**	**-0.24**	**0.03**	0.00	**0.03**	**-0.06**	-0.01
% of Scheduled Messages Before Engagement	**-0.34**	**-0.11**	**-0.06**	**-0.06**	1.00	**0.21**	**-0.54**	**0.05**	**0.07**	**0.03**	**-0.11**	**-0.04**	0.01	**-0.03**	-0.01	0.02
% of Scheduled Messages Before Disengagement	**0.46**	**0.07**	**-0.11**	**-0.11**	**0.21**	1.00	**0.71**	**0.11**	**0.07**	**-0.04**	0.00	**0.12**	**0.07**	-0.02	**-0.05**	**-0.06**
% of Scheduled Messages Engaged	**0.66**	**0.14**	**-0.05**	**-0.04**	**-0.54**	**0.71**	1.00	**0.06**	0.01	**-0.06**	**0.07**	**0.13**	**0.05**	0.00	**-0.04**	**-0.07**
Response Time to Scheduled Messages (Hours)	0.03	0.02	**0.04**	0.02	**0.05**	**0.11**	**0.06**	1.00	**0.88**	**0.09**	**-0.10**	**0.04**	-0.01	0.00	**-0.07**	**0.04**
Response Time to Any Advising Messages (Hours)	0.00	**-0.09**	0.02	0.00	**0.07**	**0.07**	0.01	**0.88**	1.00	**0.08**	**-0.11**	0.02	0.02	**-0.03**	**-0.04**	**0.04**
Proportion of Messages Asking for Help	**-0.07**	0.02	**0.36**	**0.34**	**0.03**	**-0.04**	**-0.06**	**0.09**	**0.08**	1.00	**-0.34**	0.00	0.02	**0.12**	**-0.11**	**0.04**
Average Sentiment of Replies	**0.06**	-0.02	**-0.21**	**-0.24**	**-0.11**	0.00	**0.07**	**-0.10**	**-0.11**	**-0.34**	1.00	**-0.05**	**-0.10**	**-0.11**	**0.14**	**0.08**
% of Replies About Course Planning	**0.09**	**0.08**	**0.06**	**0.03**	**-0.04**	**0.12**	**0.13**	**0.04**	0.02	0.00	**-0.05**	1.00	-0.02	**-0.22**	**-0.19**	**-0.34**
% of Replies About Financial Aid	**0.03**	0.01	0.02	0.00	0.01	**0.07**	**0.05**	-0.01	0.02	0.02	**-0.10**	-0.02	1.00	**-0.23**	**-0.06**	**-0.20**
% of Replies About Academic Planning	-0.01	**0.08**	**0.11**	**0.03**	**-0.03**	-0.02	0.00	0.00	**-0.03**	**0.12**	**-0.11**	**-0.22**	**-0.23**	1.00	**-0.17**	**-0.41**
% of Replies About Academic Supports	0.00	**-0.16**	**-0.07**	**-0.06**	-0.01	**-0.05**	**-0.04**	**-0.07**	**-0.04**	**-0.11**	**0.14**	**-0.19**	**-0.06**	**-0.17**	1.00	**-0.27**
% of Replies About Meeting Logistics	**-0.07**	0.01	**-0.08**	-0.01	0.02	**-0.06**	**-0.07**	**0.04**	**0.04**	**0.04**	**0.08**	**-0.34**	**-0.20**	**-0.41**	**-0.27**	1.00

Note: Each cell represents the Pearson's correlation coefficient between the two engagement measures indicated. Bolded numbers are significant at the p<0.05 level.

We call attention to a few surprising and noteworthy dynamics for concision. To begin with one example, we see that the percent of scheduled messages students responded to is at most weakly correlated with every other measure except for the percent of messages before engagement (-0.34) and disengagement (0.45) and the percent of scheduled messages engaged (0.61). Each of these strong relationships make mechanical sense, in that a student must have been engaged for a higher percentage of the messaging duration if they responded to a greater proportion of scheduled messages in general, and thus were more likely to have engaged earlier or disengaged later in the intervention. The general lack of correlation otherwise also indicates that response rates do not tell us much about response *content* at all, re-emphasizing the usefulness of examining engagement beyond just response rates.

Reply length (specified either as average reply length or as the proportion of substantive replies) is positively related to the proportion of messages asking for help (0.36 or 0.34) and negatively related to the average sentiment of messages (-0.21 or -0.24). This makes some intuitive sense, in that students who are asking for assistance with something would likely provide more detailed messages versus a student who has no need for assistance. That sentiment itself is also negatively related to help-asking (-0.34) also makes intuitive sense given that bids for help are often predicated on students sharing their issues (e.g., course registration portals not working properly) or hardship (e.g., inability to pay for tuition).

Note here that the response time measures seem only weakly correlated with other measures across the board, with no correlation coefficient higher than an absolute value of 0.11 (except for the coefficient between the two response speed measures, which is to be expected mechanically). This indicates that response speed seems to capture a completely different dimension of student engagement behavior than the other measures, though we will examine whether this group of measures seem to provide any worthwhile information with respect to student outcomes in the next section.

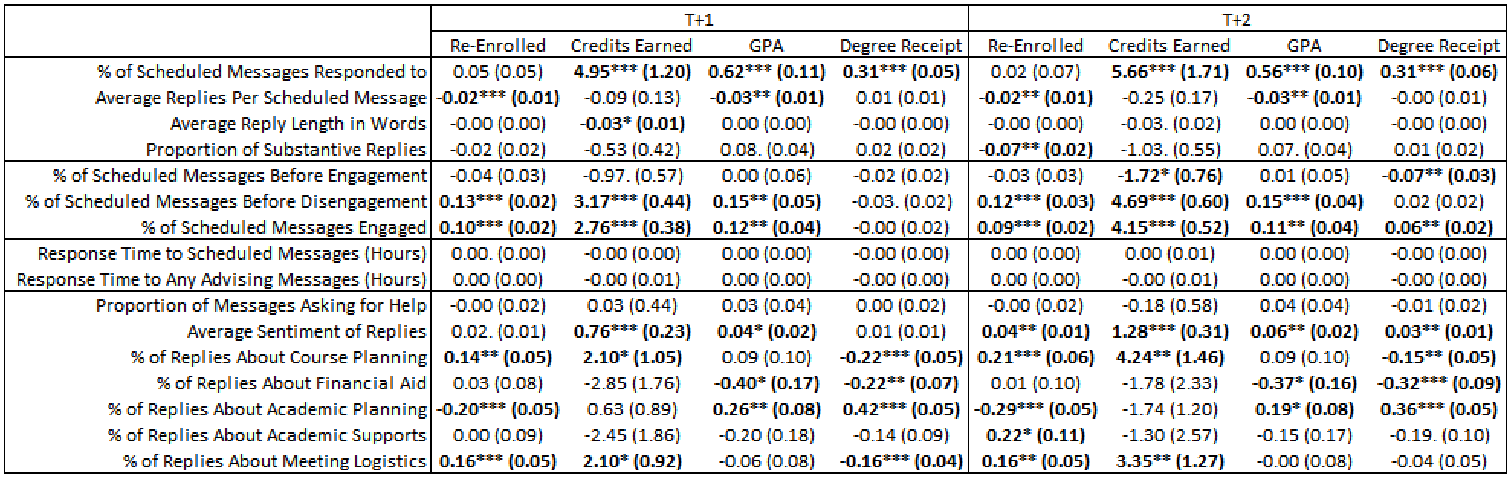
Interestingly, the topical content measures only seem related to one another, and this is largely the result of a mechanical relationship whereby the topical content measures must sum to 1 as proportions of the student replies. This again emphasizes the intuition that *what* students engage about is a critical piece of the puzzle in understanding *how* students engage with text-based advising, distinct from response rates. That said, we observe negligible relationships between topical content measures only in the case of the prevalence of financial aid against course planning (-0.02) and academic support (-0.06). This might be due to there being a higher likelihood of students discussing *both* finances and course-taking or finances and academic performance (e.g., maintaining GPAs for scholarships, or needing more academic support if finances are an issue), resulting in less negative correlations than those we observe between other topics. We also observe particularly negative correlations between meeting logistics and both academic planning and course planning. This may reflect the idea that these topics are often sufficiently complex that students and advisors would prefer to discuss them in-person rather than over text, and so we wouldn't observe the ensuing conversations about academic or course planning in our measures.

In general, we do not see particularly strong correlations in measures *across* categories, and strong correlations *within* categories tend to be mechanical in nature (e.g., topical content proportions). This again indicates that these measures seem to be capturing quite different information from one another, suggesting the potential value of thinking about student engagement along multiple dimensions when possible, beyond simple response rates.

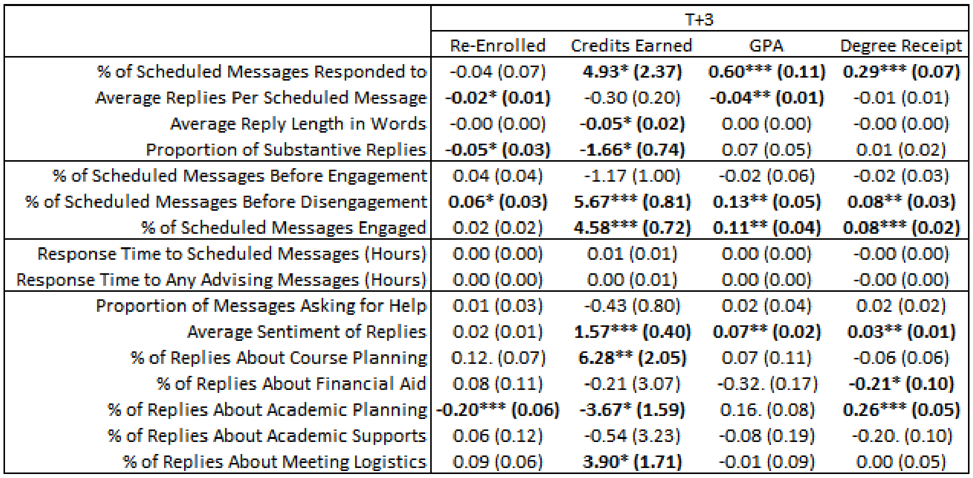
Vc. Relationships Between Engagement Measures and Student Outcomes

While the aforementioned analyses provide excellent insight into how students navigated the texting intervention, we now move to examine the extent to which these engagement measures are actually related to the outcomes of interest for the intervention. Again, we view these descriptive analyses as purely exploratory for the sake of generating hypotheses about what kinds of engagement may relate to better student outcomes, and thus what kinds of engagement future intervention designers may wish to elicit in their construction of similar text-based advising programs. Table 5 thus displays the results of many regressions (16 engagement measures by 12 outcomes), where each cell represents a separate regression as specified in Section IVe. Bolded cells indicate relationships significant at the p<0.05 level.

Table 5. Relationships Between Engagement Measures and Student Outcomes



	T+1				T+2			
	Re-Enrolled	Credits Earned	GPA	Degree Receipt	Re-Enrolled	Credits Earned	GPA	Degree Receipt
% of Scheduled Messages Responded to	0.05 (0.05)	**4.95*** (1.20)**	**0.62*** (0.11)**	**0.31*** (0.05)**	0.02 (0.07)	**5.66*** (1.71)**	**0.56*** (0.10)**	**0.31*** (0.06)**
Average Replies Per Scheduled Message	**-0.02*** (0.01)**	-0.09 (0.13)	**-0.03** (0.01)**	0.01 (0.01)	**-0.02** (0.01)**	-0.25 (0.17)	**-0.03** (0.01)**	-0.00 (0.01)
Average Reply Length in Words	-0.00 (0.00)	**-0.03* (0.01)**	0.00 (0.00)	-0.00 (0.00)	-0.00 (0.00)	-0.03. (0.02)	0.00 (0.00)	-0.00 (0.00)
Proportion of Substantive Replies	-0.02 (0.02)	-0.53 (0.42)	0.08. (0.04)	0.02 (0.02)	**-0.07** (0.02)**	-1.03. (0.55)	0.07. (0.04)	0.01 (0.02)
% of Scheduled Messages Before Engagement	-0.04 (0.03)	-0.97. (0.57)	0.00 (0.06)	-0.02 (0.02)	-0.03 (0.03)	**-1.72* (0.76)**	0.01 (0.05)	**-0.07** (0.03)**
% of Scheduled Messages Before Disengagement	**0.13*** (0.02)**	**3.17*** (0.44)**	**0.15** (0.05)**	-0.03. (0.02)	**0.12*** (0.03)**	**4.69*** (0.60)**	**0.15*** (0.04)**	0.02 (0.02)
% of Scheduled Messages Engaged	**0.10*** (0.02)**	**2.76*** (0.38)**	**0.12** (0.04)**	-0.00 (0.02)	**0.09*** (0.02)**	**4.15*** (0.52)**	**0.11** (0.04)**	**0.06** (0.02)**
Response Time to Scheduled Messages (Hours)	0.00. (0.00)	-0.00 (0.00)	0.00 (0.00)	-0.00 (0.00)	0.00 (0.00)	0.00 (0.01)	0.00 (0.00)	-0.00 (0.00)
Response Time to Any Advising Messages (Hours)	0.00 (0.00)	-0.00 (0.01)	0.00 (0.00)	-0.00 (0.00)	0.00 (0.00)	-0.00 (0.01)	0.00 (0.00)	-0.00 (0.00)
Proportion of Messages Asking for Help	-0.00 (0.02)	0.03 (0.44)	0.03 (0.04)	0.00 (0.02)	-0.00 (0.02)	-0.18 (0.58)	0.04 (0.04)	-0.01 (0.02)
Average Sentiment of Replies	0.02. (0.01)	**0.76*** (0.23)**	**0.04* (0.02)**	0.01 (0.01)	**0.04** (0.01)**	**1.28*** (0.31)**	**0.06** (0.02)**	**0.03** (0.01)**
% of Replies About Course Planning	**0.14** (0.05)**	**2.10* (1.05)**	0.09 (0.10)	**-0.22*** (0.05)**	**0.21*** (0.06)**	**4.24** (1.46)**	0.09 (0.10)	**-0.15** (0.05)**
% of Replies About Financial Aid	0.03 (0.08)	-2.85 (1.76)	**-0.40* (0.17)**	**-0.22** (0.07)**	0.01 (0.10)	-1.78 (2.33)	**-0.37* (0.16)**	**-0.32*** (0.09)**
% of Replies About Academic Planning	**-0.20*** (0.05)**	0.63 (0.89)	**0.26** (0.08)**	**0.42*** (0.05)**	**-0.29*** (0.05)**	-1.74 (1.20)	**0.19* (0.08)**	**0.36*** (0.05)**
% of Replies About Academic Supports	0.00 (0.09)	-2.45 (1.86)	-0.20 (0.18)	-0.14 (0.09)	**0.22* (0.11)**	-1.30 (2.57)	-0.15 (0.17)	-0.19. (0.10)
% of Replies About Meeting Logistics	**0.16*** (0.05)**	**2.10* (0.92)**	-0.06 (0.08)	**-0.16*** (0.04)**	**0.16** (0.05)**	**3.35** (1.27)**	-0.00 (0.08)	-0.04 (0.05)



	T+3			
	Re-Enrolled	Credits Earned	GPA	Degree Receipt
% of Scheduled Messages Responded to	-0.04 (0.07)	**4.93* (2.37)**	**0.60*** (0.11)**	**0.29*** (0.07)**
Average Replies Per Scheduled Message	**-0.02* (0.01)**	-0.30 (0.20)	**-0.04** (0.01)**	-0.01 (0.01)
Average Reply Length in Words	-0.00 (0.00)	**-0.05* (0.02)**	0.00 (0.00)	-0.00 (0.00)
Proportion of Substantive Replies	**-0.05* (0.03)**	**-1.66* (0.74)**	0.07 (0.05)	0.01 (0.02)
% of Scheduled Messages Before Engagement	0.04 (0.04)	-1.17 (1.00)	-0.02 (0.06)	-0.02 (0.03)
% of Scheduled Messages Before Disengagement	**0.06* (0.03)**	**5.67*** (0.81)**	**0.13** (0.05)**	**0.08** (0.03)**
% of Scheduled Messages Engaged	0.02 (0.02)	**4.58*** (0.72)**	**0.11** (0.04)**	**0.08*** (0.02)**
Response Time to Scheduled Messages (Hours)	0.00 (0.00)	0.01 (0.01)	0.00 (0.00)	-0.00 (0.00)
Response Time to Any Advising Messages (Hours)	0.00 (0.00)	0.00 (0.01)	0.00 (0.00)	-0.00 (0.00)
Proportion of Messages Asking for Help	0.01 (0.03)	-0.43 (0.80)	0.02 (0.04)	0.02 (0.02)
Average Sentiment of Replies	0.02 (0.01)	**1.57*** (0.40)**	**0.07** (0.02)**	**0.03** (0.01)**
% of Replies About Course Planning	0.12. (0.07)	**6.28** (2.05)**	0.07 (0.11)	-0.06 (0.06)
% of Replies About Financial Aid	0.08 (0.11)	-0.21 (3.07)	-0.32. (0.17)	**-0.21* (0.10)**
% of Replies About Academic Planning	**-0.20*** (0.06)**	**-3.67* (1.59)**	0.16. (0.08)	**0.26*** (0.05)**
% of Replies About Academic Supports	0.06 (0.12)	-0.54 (3.23)	-0.08 (0.19)	-0.20. (0.10)
% of Replies About Meeting Logistics	0.09 (0.06)	**3.90* (1.71)**	-0.01 (0.09)	0.00 (0.05)

Note: All coefficients shown above are the result of a regression as described in Section IVe that includes student academic baseline and demographic characteristics, as well as student randomization block fixed effects. Each engagement measure is then included in the regression with the given outcome of interest *without* any other engagement measure. Thus, each cell represents its own separate regression. Standard errors in parentheses. Bolded coefficients are significant at the p<0.05 level. (. = p<0.10) (* = p<0.05) (** = p<0.01) (*** = p<0.001)

Among the **Frequency and Intensity of Student Replies** measures, the percent of scheduled messages students responded to seemed to be the only measure with meaningful relationships to student outcomes: students who responded to a greater share of scheduled messages experienced substantially higher term credits, higher term GPAs, and higher likelihood of degree receipt in every term, even after controlling for academic baseline covariates. Interestingly, a higher proportion of substantive replies seems negatively correlated to re-enrollment and credit receipt in the later terms. While this could be a result of the dynamic we hypothesized earlier that more substantial messages were reflective of greater individual struggles, we surprisingly do not see the same relationship for help-asking messages.

Looking at the **Engagement Duration** measures, we observe many strong relationships all in the direction of more positive outcomes for students who were engaged for a greater share of the intervention. That is, students who engaged earlier, disengaged later, and were engaged for a larger proportion of the intervention had substantially higher re-enrollment rates, credit accumulation, and

GPA. These same students had higher levels of degree receipt, but only at the later time intervals of T+2 and T+3 with increasing magnitude further from the intervention start term.

We observe no significant relationships across the board for the **Response Speed** measures, as well as the proportion of messages asking for help. Turning to the remainder of the**Response Content** measures, higher sentiment levels (i.e., positive sentiment) correlate with slightly higher levels of credits, GPA, and degree receipt, with increasing magnitude in more distal term periods. It should be noted that these magnitudes are smaller than the other measures at least in part because it is one of the only non-proportion measures we constructed with a range of -2 to 2. Thus, a unit change in the average sentiment is more feasible in reality than a unit change in a proportion variable like the percent of scheduled messages responded to (i.e., going from 0% of scheduled messages responded to, to 100%).

For the topical content measures, increased levels of course planning were positively correlated with re-enrollment and credits earned, again with increasing magnitude over time. It was also negatively correlated with degree receipt in the earlier terms, which makes some intuitive sense given that students focused on course enrollment for the coming term are likely not ready to graduate for the given term. Discussion of academic support and meeting logistics reflected some of these same trends with student outcomes, likely for the same reasons: students looking for either of these types of support from their advisors were unlikely to be immediately ready to graduate, but seemed to benefit in terms of re-enrollment and successful completion of credits in the given term. Greater shares of discussion about financial aid was negatively correlated with both GPA and degree receipt, perhaps reflecting that academic difficulty has a strong relationship in general with student finances. The share of discussion about academic planning has, perhaps intuitively, almost the opposite relationship with outcomes to course planning. That is, higher rates of academic planning discussion was associated with far higher levels of degree receipt and far lower levels of re-enrollment and credit accumulation, likely because academic planning discussion includes topics like job applications, graduation logistics, and so on.

VI. Discussion

Taken altogether, our results suggest several high-level insights about student engagement patterns in the context of text-based advising interventions. First, we see that even interventions designed to elicit strong engagement from students don't necessarily succeed in doing so across the board. Moreover, we see that response rates alone are likely insufficient to characterize the nuance and multi-dimensionality inherent in how students engage in these personalizable interventions. We see broad variation across students in many of the four categories and sixteen measures we constructed, and these measures generally seem to capture quite different information about students' engagement from one another given low between-measure correlations. In general, we view these widely varying engagement measures as more likely to be malleable through thoughtful program design, but this will need to be explored in greater detail in future study. There is also likely a relationship between the simplicity of the engagement measure and malleability - for example, it may be relatively easy to change the average length of student response by shifting from sending students closed-response prompts (e.g., "Are you planning to submit the FAFSA?") to sending students open-response prompts (e.g., "How can I help you submit the FAFSA?"). In contrast, more complex measures such as the sentiment that students convey in their messages may prove harder (and potentially undesirable) to manipulate.

Second, we see that many of these engagement measures have statistically and substantively significant relationships with academic outcomes of interest like student persistence and degree receipt, even after holding constant students' demographics and baseline academic characteristics. Most notably, among students who texted into the campaign at least once, responding to a greater share of scheduled messages was positively correlated with better academic performance and degree completion. Similarly, *longer* periods of engagement were moreover associated with higher persistence and academic performance. Although not causal, these findings together suggest that we stand to learn much more about improving the efficacy of two-way text advising campaigns, for example by experimentally exploring how more sustained engagements with students could enhance outcomes.

In contrast, we found that longer student messages to advisors are associated with higher rates of help-seeking language and negative sentiment, but lower rates of persistence and credit completion. While it makes intuitive sense that students who seek help are typing out longer messages and expressing negative affect (e.g., frustration), that we observe worse academic performance among these students merits further exploration. For instance, it is possible that at-risk students who engage in help-seeking behaviors via text are not receiving the support they seek or need. A deeper understanding of the relationships observed here--between help-seeking language, negative affect, subsequent interactions with campus supports, and worse academic performance--could help shed light on ways to design advising programs in a way that delivers enhanced support for students who self-identify as needing help.

Additionally, we found that a greater share of student replies about meeting logistics is positively correlated with re-enrollment and credit accumulation during the first two semesters of the texting intervention. This supports the notion that one of the ways in which text-based outreach could help students is to make scheduling advising appointments easier. Particularly as the challenges that students face become increasingly complex (e.g., financial aid issues, uncertainty about plans for transferring to a four-year university), it stands to reason that text-based engagement will be useful insofar as it allows students to plan when they will meet with an advisor for more in-depth assistance.

Overall, this study calls for a more careful look under the surface of text-based advising programs such as the N2FL intervention. Our exploratory findings confirm our hypothesis that there is meaningful variation across students in terms of how they respond topically and length-wise and for how long they choose to engage in interactive texting campaigns spanning multiple academic semesters. A deeper understanding of the heterogeneity in student engagement behaviors and the identification of specific behaviors that are correlated with academic success could help scholars and practitioners alike design text-based advising programs with greater intentionality, precision, and efficacy.

Usefully, all the measures we construct are imminently scalable and applicable to any similar texting context, meaning that they can serve as more consistent tools to help us better understand and contextualize the results of interventions that have previously taken place, as well as diagnostics to inform program management and implementation *as an intervention is happening*. To push for greater codification of such interaction measures across intervention contexts, we also offer our code open-source for other researchers to build upon and implement in their own studies. This ability to perform real-time diagnostics is especially appealing in that local institutions can glean important insights about their specific student population (that might not be applicable in other contexts) and test approaches to adjust their messaging strategy accordingly. With the combination of these more nuanced standard engagement measures and sophisticated NLP techniques, we can offer researchers and practitioners greater visibility into important dynamics like student uptake going forward.

VII. References

Airoldi, E. M., & Bischof, J. M. (2016). Improving and Evaluating Topic Models and Other Models of Text. *Journal of the American Statistical Association*, *111*(516), 1381–1403. https://doi.org/10.1080/01621459.2015.1051182

Alikaniotis, D., & Raheja, V. (2019, August 7). Under the Hood at Grammarly: Leveraging Transformer Language Models for Grammatical Error Correction | . *Grammarly Engineering Blog*. https://www.grammarly.com/blog/engineering/under-the-hood-at-grammarly-leveraging-transformer-language-models-for-grammatical-error-correction/

Ambartsoumian, A., & Popowich, F. (2018). Self-Attention: A Better Building Block for Sentiment Analysis Neural Network Classifiers. *ArXiv:1812.07860 [Cs]*. https://doi.org/10.18653/v1/P17

Arnold, K. D., Owen, L., & Lewis, J. (2020). Inside the Black Box of Text-Message College Advising. *Journal of College Access*, *5*(2), 5.

Avery, C., & Turner, S. (2012). Student Loans: Do College Students Borrow Too Much – Or Not Enough? *Journal of Economic Perspectives* 26(1), 165-193.

Avery, C., Castleman, B. L., Hurwitz, M., Long, B. T., & Page, L. C. (2021). Digital messaging to improve college enrollment and success. National Bureau of Economic Research Working Paper No 27897. Retrieved from https://www.nber.org/system/files/working_papers/w27897/w27897.pdf

Barbieri, F., Camacho-Collados, J., Neves, L., & Espinosa-Anke, L. (2020). TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. *ArXiv:2010.12421 [Cs]*. http://arxiv.org/abs/2010.12421

Bettinger, E. P., Castleman, B. L., Choe, A., & Mabel, Z. (2021). Finishing the Last Lap: Experimental Evidence on Strategies to Increase College Completion for Students At Risk of Late Withdrawal. EdWorkingPaper No. 21-488. Retrieved from https://edworkingpapers.org/sites/default/files/ai21-488.pdf

Bird, K. A., Castleman, B. L., Denning, J. T., Goodman, J., Lamberton, C., & Rosinger, K. O. (2021). Nudging at scale: Experimental evidence from FAFSA completion campaigns. *Journal of Economic Behavior & Organization*, *183*, 105-128.

Blei, D. M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3,* 30.

Bound, J., Lovenheim, M., & Turner, S. (2010). Why have college completion rates declined? An analysis of changing student preparation and collegiate resources. *American Economic Journal, 2*(3), 129-157.

Carnevale, A. P., Jayasundera, T., & Gulish, A. (2016). *America's Divided Recovery: College Haves and Have-Nots*. Georgetown University Center on Education and the Workforce.

Cartujano-Barrera, F., Arana-Chicas, E., Ramírez-Mantilla, M., Perales, J., Cox, L. S., Ellerbeck, E. F., ... & Cupertino, A. P. (2019). "Every day I think about your messages": assessing text messaging engagement among Latino smokers in a mobile cessation program. *Patient preference and adherence*, *13*, 1213.

Castleman, B.L. & Page, L. (2016). Freshman year financial aid nudges: An experiment to increase FAFSA renewal and college persistence. *Journal of Human Resources, 51*(2), 389-415.

Castleman, B. L., Meyer, K. E., Sullivan, Z., Hartog, W. D., Miller, S. (2017). Nudging students beyond the FAFSA: The impact of university outreach on financial aid behaviors and outcomes. *Journal of Student Financial Aid*, 47(3) 2.

Denning, J., Eide, E., & Warnick, M. (2019). Why have college completion rates increased? Working Paper No. 12411. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3408309

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

Duppada, V., Jain, R., & Hiray, S. (2018). SeerNet at SemEval-2018 Task 1: Domain Adaptation for Affect in Tweets. *ArXiv:1804.06137 [Cs]*. http://arxiv.org/abs/1804.06137

Fesler, L. (2020). Opening the Black Box of College Counseling. CEPA Working Paper No. 20-03. Retrieved from https://files.eric.ed.gov/fulltext/ED605975.pdf.

Fesler, L., Dee, T., Baker, R., & Evans, B. (2019). Text as Data Methods for Education Research. *Journal of Research on Educational Effectiveness*, *12*(4), 707–727. https://doi.org/10.1080/19345747.2019.1634168

Gurantz, O., Pender, M., Mabel, Z., Larson, C., & Bettinger, E. (2020). Virtual advising for high-achieving high school students. Economics of Education Review, 75, 101974.

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Gopalan, M., & Brady, S. (2019). College students' sense of belonging: A national perspective. *Educational Researcher, 49*(2).

Irvine, L., Melson, A. J., Williams, B., Sniehotta, F. F., McKenzie, A., Jones, C., & Crombie, I. K. (2017). Real time monitoring of engagement with a text message intervention to reduce binge drinking among men living in socially disadvantaged areas of Scotland.*International journal of behavioral medicine*, *24*(5), 713-721.

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. *ArXiv:1408.5882 [Cs]*. http://arxiv.org/abs/1408.5882

Kraft, M. A., & Dougherty, S. M. (2013). The effect of teacher–family communication on student engagement: Evidence from a randomized field experiment. *Journal of Research on Educational Effectiveness*, *6*(3), 199-222.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations.*ArXiv:1909.11942 [Cs]*. http://arxiv.org/abs/1909.11942

Mabel, Z., Castleman, B., & Bettinger, E. (2019). Finishing the last lap: Experimental evidence on strategies to increase college completion for students at risk of late departure. Working Paper. Retrieved from https://scholar.harvard.edu/zmabel/publications/finishing-last-lap-experimental-evidence-strategies-increase-college-completion

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing Semantic Coherence in Topic Models. *Proceedings of the Conferenceon Empirical Methods in Natural Language Processing*, 262–272. http://dl.acm.org/citation.cfm?id=2145432.2145462

Mohammad, S., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S. (2018). SemEval-2018 Task 1: Affect in Tweets. *Proceedings of The 12th InternationalWorkshop on Semantic Evaluation*, 1–17. https://doi.org/10.18653/v1/S18-1001

Morris, J. X., Lifland, E., Yoo, J. Y., Grigsby, J., Jin, D., & Qi, Y. (2020). TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. *ArXiv:2005.05909 [Cs]*. http://arxiv.org/abs/2005.05909

Nelson, L. A., Spieker, A., Greevy, R., LeStourgeon, L. M., Wallston, K. A., & Mayberry, L. S. (2020). User Engagement Among Diverse Adults in a 12-Month Text Message–Delivered Diabetes Support Intervention: Results from a Randomized Controlled Trial. *JMIR mHealth and uHealth*, *8*(7), e17534.

O'Brien, H.L. & Toms, E.G. (2008). What is user engagement? A conceptual framework for defining user engagement with technology. Journal of the American Society for Information Science & Technology, 59(6), 938- 955. DOI: 10.1002/asi.20801.

Oreopoulos, P. & Petronijevic, U. (2018). Student coaching: How far can technology go? *Journal of Human Resources, 53*(2), 299-329.

Page, L. C., & Gehlbach, H. (2017). How an artificially intelligent virtual assistant helps students navigate the road to college. *AERA Open*, *3*(4), 1-12.

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1–135. https://doi.org/10.1561/1500000011

Park, J. H., Shin, J., & Fung, P. (2018). Reducing Gender Bias in Abusive Language Detection. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* 2799–2804. https://doi.org/10.18653/v1/D18-1302

Penner, E. K., Rochmes, J., Liu, J., Solanki, S. M., & Loeb, S. (2019). Differing Views of Equity: How Prospective Educators Perceive Their Role in Closing Achievement Gaps. *The Russell Sage Foundation Journal of the Social Sciences*, *5*(3), 103–127. https://doi.org/10.7758/RSF.2019.5.3.06

Psihogios, A. M., Li, Y., Butler, E., Hamilton, J., Daniel, L. C., Barakat, L. P., ... & Schwartz, L. A. (2019). Text message responsivity in a 2-way short message service pilot intervention with adolescent and young adult survivors of cancer. *JMIR mHealth and uHealth*, *7*(4), e12547.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *ArXiv:1910.10683 [Cs, Stat]*. http://arxiv.org/abs/1910.10683

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). stm: An R Package for Structural Topic Models. *Journal of Statistical Software*, *91*(2). https://doi.org/10.18637/jss.v091.i02

Shapiro, D., Ryu, M., Huie, F., & Liu, Q. (October 2019). Some College, No Degree , a 2019 Snapshot for the Nation and 50 States, Signature Report No. 17, Herdon, VA: National Student Clearinghouse Research Center

Smythe-Leistico, K., & Page, L. C. (2018). Connect-text: Leveraging text-message communication to mitigate chronic absenteeism and improve parental engagement in the earliest years of schooling. *Journal of Education for Students Placedat Risk (JESPAR)*, *23*(1-2), 139-152.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642. https://www.aclweb.org/anthology/D13-1170

Sullivan, Z., Castleman, B., & Bettinger, E. (2019). College advising at a national scale: Experimental evidence from the CollegePoint initiative. EdWorkingPaper No. 19-123. Retrieved from https://edworkingpapers.com/index.php/ai19-123

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., & Wang, W. Y. (2019). Mitigating Gender Bias in Natural Language Processing: Literature Review. *ArXiv:1906.08976 [Cs]*. http://arxiv.org/abs/1906.08976

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Languageand Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

van Aken, B., Winter, B., Löser, A., & Gers, F. A. (2019). How Does BERT Answer Questions? A Layer-Wise Analysis of Transformer Representations. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 1823–1832. https://doi.org/10.1145/3357384.3358028

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *ArXiv:1706.03762 [Cs]*. http://arxiv.org/abs/1706.03762

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2020). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *ArXiv:1906.08237 [Cs]*. http://arxiv.org/abs/1906.08237

Zhang, S., Hamburger, E., Kahanda, S., Lyttle, M., Williams, R., & Jaser, S. S. (2018). Engagement with a text-messaging intervention improves adherence in adolescents with type 1 diabetes: brief report. *Diabetes technology & therapeutics*, *20*(5), 386-389.

VIII. Appendix

VIIIa. Advising Models

Table A1 illustrates the four primary advising models that emerged in terms of how institutions staffed the text messaging campaign.

Table A1. Advising Models Used in the N2FL Intervention

Model	**Example Advisor Background(s)**	**Advisor Role**	**Sample Message**	**Number of Institutions**
Professional Advisor	Hired specifically for the N2FL project	Direct assistance with tasks (e.g., registering for courses, financial aid applications)	Hi, it's <Professional Advisor>. With finals coming up, I wanted to check if you've used <Support Center> for help with classes. Can I help you get connected?	Nine
Faculty Advisor	University faculty	Direct assistance with questions in their specialization (e.g., course selection) and recommending campus resources for other questions (e.g., financial aid)	Hey, it's <Faculty Advisor>. As you're planning for spring, think about picking up an extra course. This can help you graduate sooner. Can I help you choose another class?	One
Staff Point Person	Administrative assistant on student engagement team	Direct students to the resource most appropriate for providing assistance	Hi <Student>! Registration for fall and summer starts 4/2. Have you talked to an advisor about the next classes you need to take in your program?	Six
Segmented Advising	Mix of campus staff (e.g., some faculty advisors coupled	Leveraged multiple staff depending on	Hi, it's <Advisor>. Fafsa.gov is now open for the	Two

	with a career services counselor)	question (e.g., student replies to automated questions about course registration went to an Academic Advisor's portfolio)	2018-2019 school year and applying early gets you the most financial aid. Have you started FAFSA yet? [student replies are routed to a Financial Advisor's inbox]	

VIIIb. Sentiment Analysis Introduction and Methodology

Sentiment analysis is a common NLP task in which analysts use an algorithm to "read" a given text string and rate the string as containing/expressing positive, negative, or neutral sentiment (Pang & Lee, 2008). This task is especially common in commercial applications (e.g. analyzing consumer sentiment towards your product by analyzing tweets) but is becoming more pervasive in the field of education research as well (Fesler et al., 2019). One can think of this process as generating output similar to hand-coded qualitative analysis, but in an automated and highly scalable way that facilitates quantitative analysis.

There are a wide variety of techniques that data scientists use in this pursuit, but the recent NLP literature has coalesced around complicated neural network algorithms known as "transformers" (Vaswani et al., 2017). These transformer algorithms perform substantially better than previous sentiment analysis approaches because the transformer's specific architecture allows it to better account for the complexities of word context (e.g. that "I wish I were happy" actually indicates sadness), multiple word meanings (e.g. that "bank" has two separate meanings in "The river bank was wet" and "I went to the bank this morning"), and informalities (e.g. "that was sick, dude;" Ambartsoumian & Popowich, 2018).

In brief, a transformer neural network is a neural network algorithm that has been fed immense volumes of text data (such as aggregated Wikipedia articles, novels, and news articles) to generate a nuanced statistical model describing how words are put together into sentences - called a "language model." This can be thought of as giving the algorithm a generalized understanding of grammar, syntax, vocabulary, and word relationships by example. For instance, it will have seen thousands of examples of "...it is hot outside…" in its training data, but likely no examples of "...outside it hot is..." nor "...clam hot it outside...", teaching it what combinations and sequences of words are considered valid. Despite the bluntness of this approach, it is so effective at capturing complex idiosyncrasies within language that it now drives some of the most advanced and widely-used grammar checking engines (e.g. Grammarly; Alikaniotis & Raheja, 2019).

Once that language modeling process is complete, analysts then "fine-tune" the algorithm to perform a more specific task, such as sentiment analysis, using a traditional supervised machine learning framework (i.e. provide the algorithm a set of example texts with ground-truth sentiment scores so that it can optimize for accurate scoring on its own). The motivation behind separating the language modeling task from the classification task is somewhat analogous to the idea that it is easier to teach someone to play a new sport when they already have a good grasp of basic physics, fitness, and competition, versus starting from a completely blank slate. Similarly, because the transformer is well-trained in general language, it can leverage this understanding to better approach new language-based tasks afterward.

The unique contribution of the transformer architecture is a mechanism called "attention" that allows it to more effectively process longer strings of text at once by weighting words according to their functional importance in the text (e.g. using a subject introduced two paragraphs earlier to interpret a referential statement in the sentence at hand). Algorithms using the transformer architecture have literally revolutionized the landscape of NLP, pushing the state-of-the-art for model performance on nearly every single performance task and benchmark, including sentiment analysis (Devlin et al., 2019).

Choosing which transformer to use for a sentiment analysis task is a highly consequential decision. While sharing the same general principles, transformers vary due to different training datasets, different underlying mechanics and optimization processes, and different end-applications in mind. Ideally, we would be able to find a robust transformer that is trained on text data similar to ours so that we could be more assured of its appropriateness for our context.[12] Lacking that, we have opted to employ an "ensemble" approach that combines several of these transformers together. More concretely, we use five pre-existing transformer algorithms to produce sentiment analysis scores for every text message in our data, and then combine these separate classifications together in a data-driven manner using a random forest algorithm to produce a final sentiment score. By the end of this process, each text sent and received during the intervention is assigned a sentiment score from -2 to 2, corresponding with very negative, negative, neutral, positive, or very positive sentiment.

This approach is attractive because it leverages the unique strengths and insights of each of these separate models while mitigating some of their potentially problematic idiosyncrasies - the intuition here being that if each model weighs different considerations in its individual decision, they can each contribute valuably distinct insights to be incorporated into the final model. For a more detailed discussion of our constituent models, the model construction process, and performance benchmarks, please see Appendix VIIIc. In sum, we find that our ensemble model matches current state-of-the-art performance on the most common sentiment analysis benchmark, the Stanford Sentiment Treebank (SST) test (Socher et al., 2013).

We operationalize the definition of sentiment for the present study as the*perceived positivity of emotions and ideas present in a given text*. This definition then is a conglomeration of the speaker's stated emotions ("I feel sad" v. "I am excited"), communicated intention ("I hope you die" v. "I wish you the best!"), and, at least to some extent, topical content ("financial hardship" v. "vacation time"). Note that this definition is complicated for longer strings of text, in which multiple emotions/implications may be present and the overall sentiment becomes ambiguous.[13]

We argue this definition is appropriate for our context because of the nature of the N2FL text data: texts were generally only one or two sentences long (making the ambiguity of sentiment in

[12] While it would be conceptually attractive to "fine-tune" our own transformer model to best account for our educational context, training these models is both logistically and computationally complex. The authors are exploring this opportunity for related work going forward.

[13] While we would like to lean on a more standardized definition, we were unable to find a detailed and widely-accepted definition for sentiment in transformer-based models in the literature. Interestingly, sentiment as a construct across modern data science (i.e. neural network-based models rather than dictionary models) is almost entirely dependent on the SST's definition due to the strong incentive for data scientists to optimize their algorithm's SST performance for benchmarking purposes. That said, the SST intentionally encouraged their human coders to view sentiment as a flexible and subjective notion, making a formal definition elusive.

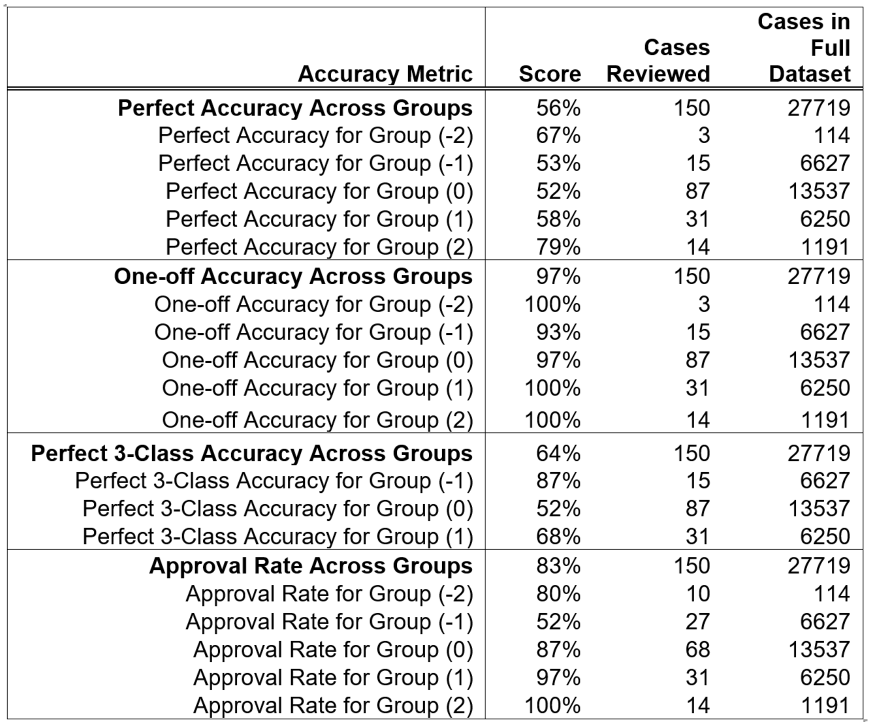
long text strings less problematic), students encounter these texts "as-is" (e.g. they are not transcribed spoken words with greater context than what we observe), and their perception of a text's sentiment is likely driven by a combination of factors (e.g. stated emotions, communicated intention, topical content).

To provide evidence for the validity of the algorithm output and its concordance with our definition of sentiment, we conduct two validation exercises:

1. Pull a random sample of N2FL texts, have human coders briefed in the construct definition manually classify each text, and then compare the human codes against the algorithm codes using traditional accuracy statistics (with human codes set as the ground-truth).
2. Pull a random sample of N2FL texts (distinct from the sample constructed in exercise 1) alongside their assigned sentiment scores from the algorithm, have human coders briefed in the construct definition approve or disapprove of each pairing's accuracy, and calculate overall and class-by-class approval rates.

A summary of the validation exercise results are displayed below in Table B1, and more details on each procedure can be found in Appendix VIIId. Note that these exercises were conducted using only one human coder for now (a coauthor on this paper); we plan to expand this process to multiple trained coders in a later draft to improve robustness.

Table B1. Validation Exercise Results



Accuracy Metric	Score	Cases Reviewed	Cases in Full Dataset
Perfect Accuracy Across Groups	56%	150	27719
Perfect Accuracy for Group (-2)	67%	3	114
Perfect Accuracy for Group (-1)	53%	15	6627
Perfect Accuracy for Group (0)	52%	87	13537
Perfect Accuracy for Group (1)	58%	31	6250
Perfect Accuracy for Group (2)	79%	14	1191
One-off Accuracy Across Groups	97%	150	27719
One-off Accuracy for Group (-2)	100%	3	114
One-off Accuracy for Group (-1)	93%	15	6627
One-off Accuracy for Group (0)	97%	87	13537
One-off Accuracy for Group (1)	100%	31	6250
One-off Accuracy for Group (2)	100%	14	1191
Perfect 3-Class Accuracy Across Groups	64%	150	27719
Perfect 3-Class Accuracy for Group (-1)	87%	15	6627
Perfect 3-Class Accuracy for Group (0)	52%	87	13537
Perfect 3-Class Accuracy for Group (1)	68%	31	6250
Approval Rate Across Groups	83%	150	27719
Approval Rate for Group (-2)	80%	10	114
Approval Rate for Group (-1)	52%	27	6627
Approval Rate for Group (0)	87%	68	13537
Approval Rate for Group (1)	97%	31	6250
Approval Rate for Group (2)	100%	14	1191

Note: Both "Cases Reviewed" and "Cases in Full Dataset" indicate the number of class cases within the exercise dataset/full dataset as labeled by the *algorithm output*. Our random sample for each exercise was stratified based on algorithm output class and message type (sent by student, scheduled text sent by advisor, personalized text sent by advisor).

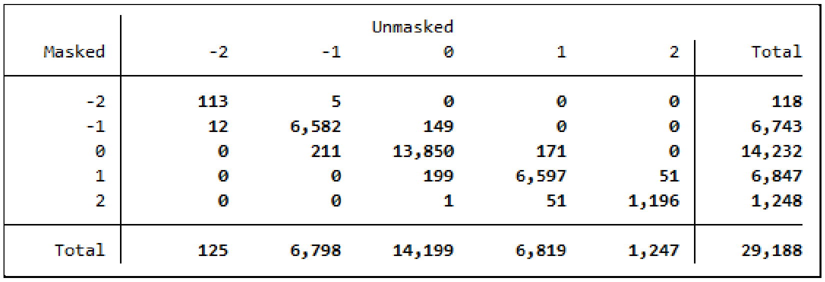
We find that our algorithm performs about as well in terms of perfect accuracy measures on the N2FL data as it does on the benchmark SST5 at 56%. This is high by sentiment analysis standards, but still suggests our analysis will suffer from measurement error. In addition, if we consider the N2FL text data a "test" dataset per supervised machine learning frameworks, the comparable performance of our algorithm on both N2FL data and the SST data may suggest: **(1)** that our algorithm was not reliant on idiosyncrasies specific to the SST data, and it is actually reading some true, generalizable signal within the text data to inform its classifications, and/or **(2)** that the N2FL text data may not be substantially different from the SST text data despite the difference in contexts and sources. This in mind, we now have evidence that using models trained on the SST rather than text data closer to the circumstances of N2FL is appropriate for our purposes.

Because the algorithm outputs a predicted probability of each possible sentiment score classification, we can also run diagnostics on the relative confidence of each of its predictions. For example, we might be more skeptical of the algorithm's sentiment classifications if it's torn between two very likely options (e.g. assigning a score of -2 with 40% probability, and a score of -1 with 38% probability) versus if it's selecting one classification with high certainty (e.g. assigning a score of 2 with 76% probability). We find in general that the algorithm is fairly confident in its classifications, and that there are few "close calls" among the N2FL texts. We dissect the results of this diagnostic test in more detail in Appendix VIIIe.

Finally, we have plentiful evidence from NLP bias research that gendered names and pronouns can systematically skew the results of text classification algorithms because these algorithms are trained on datasets that implicitly contain the biases of societal writing more broadly (Park et al., 2018; Sun et al., 2019). For example, the algorithm may interpret "She is assertive" as negative, but "He is assertive" as positive; similarly, the algorithm may interpret "Jane is assertive" as negative, but "Joe is assertive" as positive. Our text data removed names for de-identification purposes (replaced with a token stand-in, "advisorname"), making gendered names a non-issue. Moreover, because students and advisors most often spoke in the first- and second-person (I/you/we), we find exceptionally low prevalence of gendered pronouns (he/him/his/she/her/hers) in our data: out of 27,942 unique texts, only 552 (2%) contained any gendered pronoun. While we cannot rule out residual gender and racial bias in our algorithm, we have good reason to believe their impact on our analysis is negligible after these processing steps given our data context.

We also directly compare the output of the algorithm before *and* after replacing any gendered pronouns ("masking" the data) for exploratory purposes (Figure B1 below) and find that only 3% of texts change their sentiment scores at all. The overwhelming majority that do change classifications vary only slightly from masked to unmasked datasets. We intend to include more in-depth analyses of any classification inconsistencies here in a future draft.

Figure B1. Classification Concordance Table Between Masked and Unmasked Text Data



```
                             Unmasked
   Masked          -2         -1          0          1          2     Total

       -2         113          5          0          0          0       118
       -1          12      6,582        149          0          0     6,743
        0           0        211     13,850        171          0    14,232
        1           0          0        199      6,597         51     6,847
        2           0          0          1         51      1,196     1,248

    Total         125      6,798     14,199      6,819      1,247    29,188
```

VIIIc. Ensemble Model Construction

As mentioned in the prior section, our algorithm is what we refer to as an "ensemble" method. This approach involves training several models to conduct the same task, and then utilizing each of those models' output as inputs into a final model that considers each of these models' output in a final classification. This is akin to gathering a panel of experts on an issue and making a decision based on their combined recommendations. While each expert may see the same data and evidence, the variance in their interpretations may lead to importantly different conclusions worth considering.

The majority of our constituent models are built using transformer neural networks (BERT, ALBERT, XLNet, T5, RoBERTa) as described in the main narrative. In Table C1 below, we provide a rough breakdown of each of these algorithms in terms of their language model training data, task training data (i.e. sentiment analysis training data), and benchmark performance. These details are important to keep in mind as we interpret the results of our algorithm - the language model training data tells us what contexts it learned its general understanding from, and the task training data tells us what contexts it learned to classify sentiment from. For example, the BERT model we utilize was trained specifically on Yelp restaurant reviews for sentiment classification - a context where even a "lukewarm" sentence may really correspond with a quite negative sentiment score.

Table C1. Constituent Model Characteristics

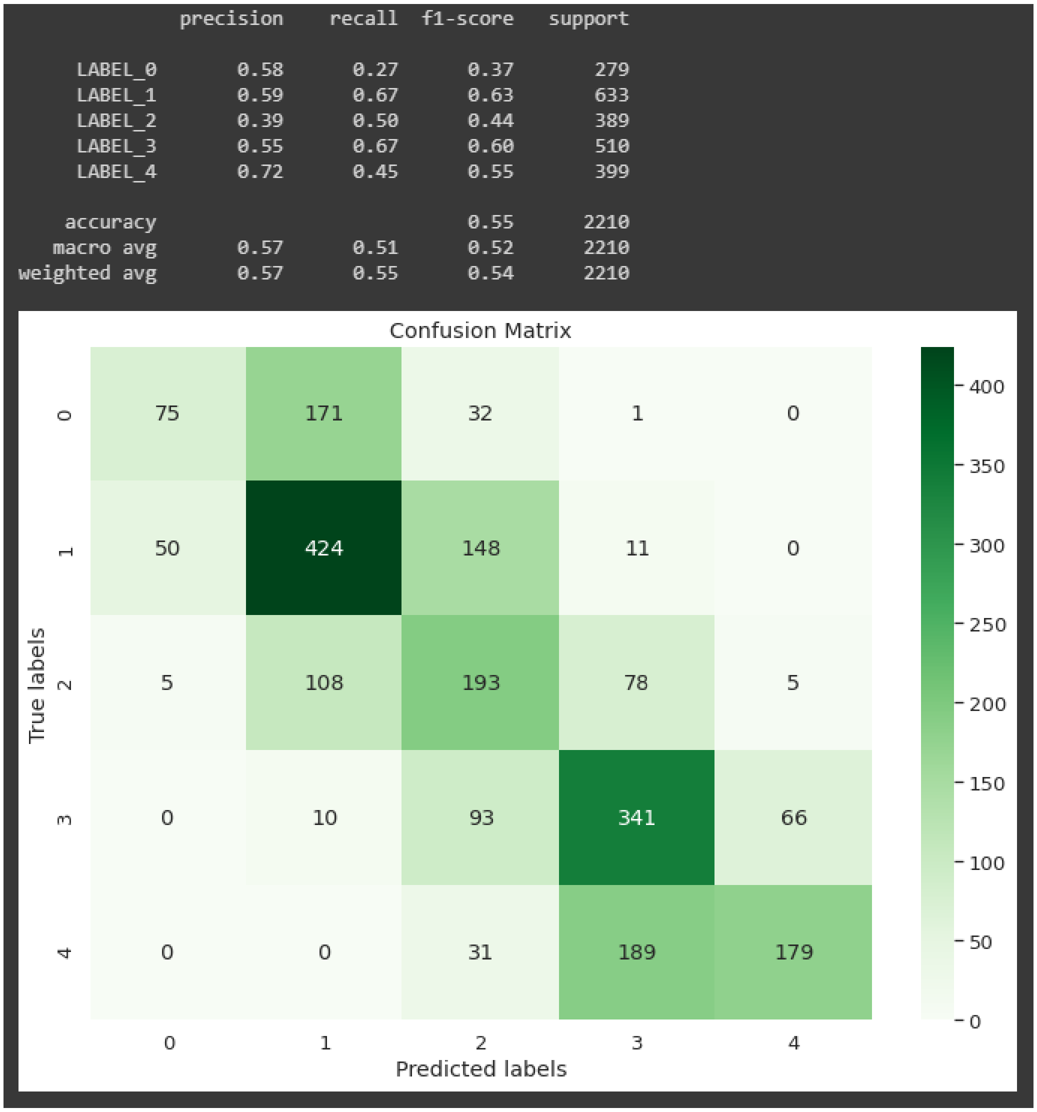
Model Name	**Language Model Training Data**	**Task Training Data**	**Sentiment Classification Type**	**Task Performance**
BERT (Google)	- 11,000 books from SmashWords ("BookCorpus") - English Wikipedia articles	Yelp Restaurant Reviews	5-class (review stars)	**SST5**: 40% Accuracy
BERT (Google)	- 11,000 books from SmashWords ("BookCorpus") - English Wikipedia articles	150,000 product reviews (1-5 stars)	5-class (review stars)	**SST5**: 42% Accuracy
ALBERT (Google)	- 11,000 books from SmashWords ("BookCorpus") - English Wikipedia articles	Movie Reviews (Stanford Sentiment Treebank, 2-class)	2-class (pos/neg)	**SST2**: 94% Accuracy

XLNet (Carnegie Mellon and Google)	- 11,000 books from SmashWords (“BookCorpus”) - English Wikipedia articles - News Articles (“Gigaword 5th Edition”) - Websites (“Common Crawl” and “ClueWeb”)	Movie Reviews (Stanford Sentiment Treebank, 2-class)	2-class (pos/neg)	**SST2**: 94% Accuracy
Stanza (Stanford)	N/A (not pre-trained)	- Movie Reviews (Stanford Sentiment Treebank, 3-class) - Sitcom Dialogue (“MELD”) - IMDB, Amazon, and Yelp reviews (UCIrvine “SLSD”) - TripAdvisor Hotel Reviews (“ArguAna”) - Tweets re: Airlines (“CrowdFlower”)	3-class (pos/neg/neu)	**SST3**: 73% Accuracy
RoBERTa (University of Washington and Facebook AI)	- 11,000 books from SmashWords (“BookCorpus”) - English Wikipedia articles - News Articles (“Common Crawl News” - Web content extracted from websites shared on Reddit (“OpenWebText”) - Story dataset (“CommonCrawl” Stories) - ~1 year of Tweets	Tweets (TweetEval, per Barbieri et. al, 2020)	3-class (pos/neg/neu)	**SST3:** 65% Accuracy
T5 (Google)	- Websites (“Clean Common Crawl”)	Highly polarized IMDB movie reviews	2-class (pos/neg)	**SST2**: 90% Accuracy

In our process, we have each algorithm classify each text and provide its calculated probabilities for each possible classification (e.g. 52% probability of a very negative sentiment, 30% of negative, 12% of neutral, 6% of positive, and 0% of very positive; these scores will always sum to 100%). We then use these outputs as inputs into the random forest classifier, which is finally trained using the Stanford Sentiment Treebank, 5-class set.

Our algorithm has a **base accuracy score** on the **SST5** of 55% (proportion of perfect classifications). This is tied for the current state-of-the-art across all NLP research to date. In Figure C1 below, I display the accuracy diagnostics of our random forest algorithm on the SST5 test set. The confusion matrix is really the key figure; if the algorithm performed perfectly, we would see all observations would fall into the diagonal cells. Note that as is common for these fine-grain sentiment analysis classifications, our model performs noticeably less well at detecting neutrality in these data.

Figure C1. Accuracy Diagnostics for Random Forest Classifier on SST5



```
              precision    recall  f1-score   support

     LABEL_0       0.58      0.27      0.37       279
     LABEL_1       0.59      0.67      0.63       633
     LABEL_2       0.39      0.50      0.44       389
     LABEL_3       0.55      0.67      0.60       510
     LABEL_4       0.72      0.45      0.55       399

    accuracy                           0.55      2210
   macro avg       0.57      0.51      0.52      2210
weighted avg       0.57      0.55      0.54      2210
```

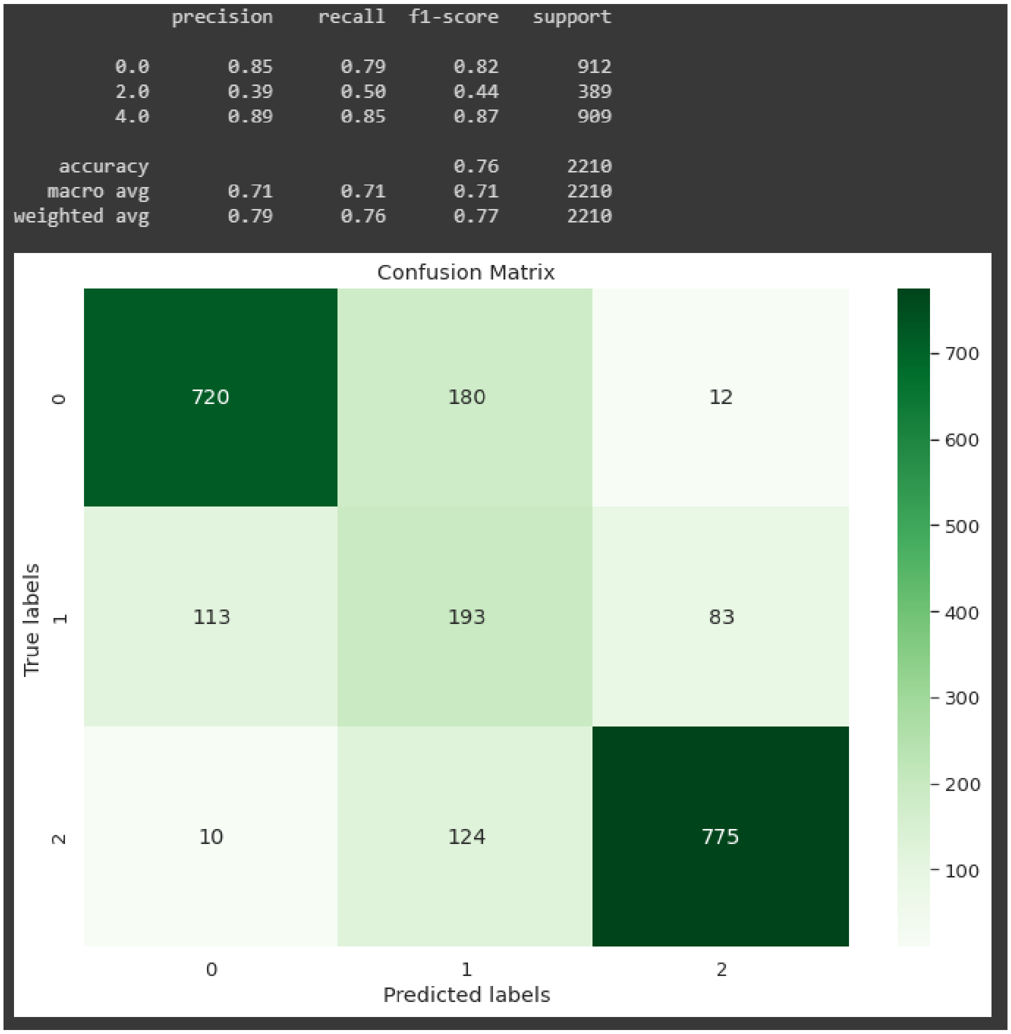
Confusion Matrix

	0	1	2	3	4
0	75	171	32	1	0
1	50	424	148	11	0
2	5	108	193	78	5
3	0	10	93	341	66
4	0	0	31	189	179

Another common way of assessing accuracy on these 5-class sentiment scores is to consider the “one-off accuracy” - or, what proportion of cases the score is only one point off of the true score. This is a good way to gauge how far off the algorithm is when it provides an incorrect classification. In our case, we have a one-off accuracy rate of 96%. In other words, even when our algorithm is wrong for the exact classification, it’s not off by much (i.e. it is not seeing a very positive text and calling it very negative, or even neutral).

Yet another way of slicing performance is by thinking only of “valence” (emotional direction) without magnitude (e.g. combine negative and very negative scores into just a single negative category). Using the standard Stanford Sentiment Treebank 3-class test, we achieve a base accuracy score of 76%. This is a less common task in the most recent wave of NLP research, and so it is unclear how this performs relative to the state of the art. For reference, Stanford’s Stanza model (which is a constituent model of ours) achieves an accuracy of 70%. Figure C2 below displays the accuracy diagnostics on the 3-class set.

Figure C2. Accuracy Diagnostics for Random Forest Classifier on SST3



```
              precision    recall  f1-score   support

         0.0       0.85      0.79      0.82       912
         2.0       0.39      0.50      0.44       389
         4.0       0.89      0.85      0.87       909

    accuracy                           0.76      2210
   macro avg       0.71      0.71      0.71      2210
weighted avg       0.79      0.76      0.77      2210
```

VIIId. Validating Algorithm Classification Output

In brief, we conduct our validation exercises according to the following procedure:

1. Comparing human-coded texts to algorithmically-coded texts
    a. Pull a random sample of 5 unique N2FL texts from each crossed group defined by the message type (scheduled text from advisor, personalized text from advisor, text from student) and algorithm sentiment score (e.g. sample from those given a score of -2, then from those given a score of -1, etc.), for a total of 75 texts. Sample an additional 75 texts completely at random, for a total of 150 texts.
    b. Brief human coders in our operationalized definition, debriefing texts shown in Table 3 in the main text. For now, we have only one human coder, but intend to expand this group to at least 5 in a future draft.
    c. Show human coders the texts in random order, **without** the algorithm sentiment scores attached, and ask them to rate each text from -2 to 2 per our sentiment definition. If they were unsure, they were asked to still provide their best guess.

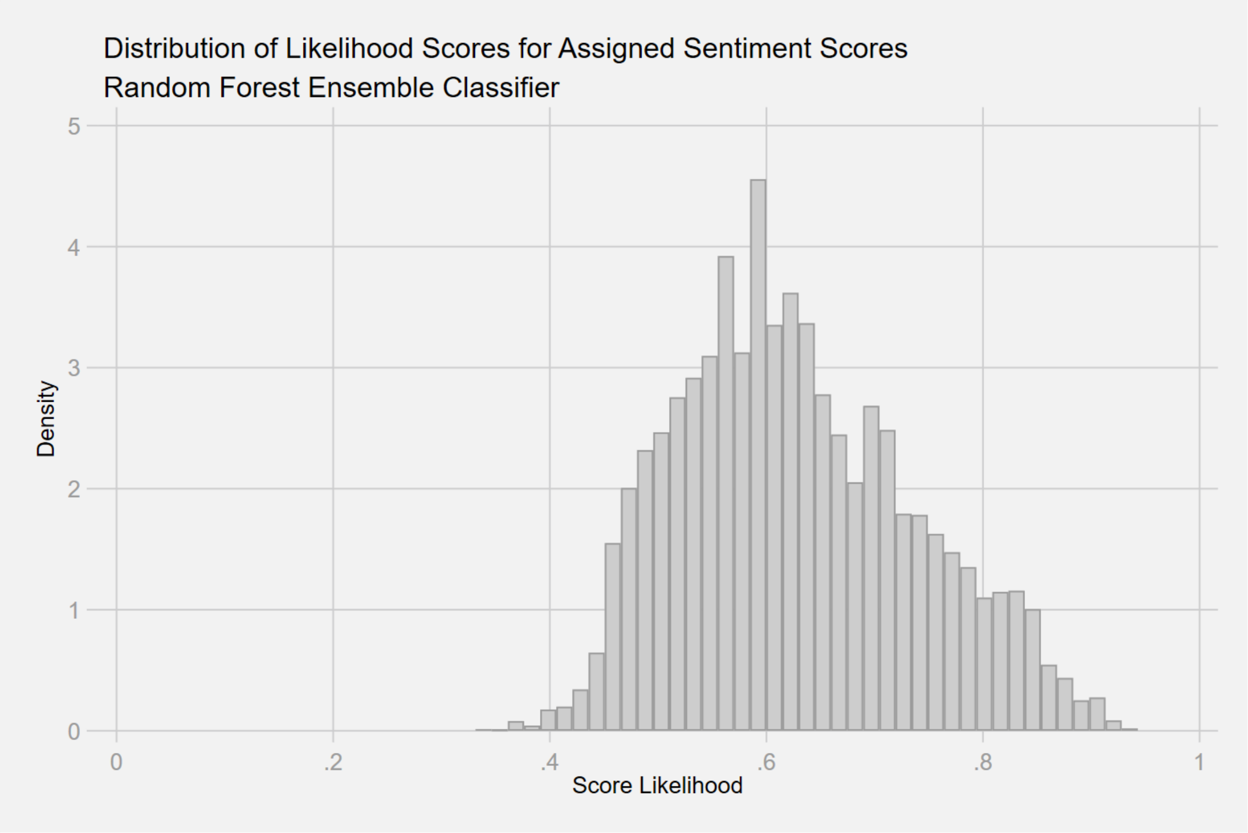
    d. Consolidate human-coded texts by taking the average score, rounded to the nearest integer value.
    e. Compare the scores given by the consensus of human-coders against the algorithm's scores overall and by each grouping of algorithm-coded sentiment score.

2. Auditing algorithmically-coded texts for human approval
    a. Create a random sample exactly as described for exercise 1, but with a different random seed (such that there may be overlap between the two samples, but they are generated totally independently)
    b. Brief human coders in our operationalized definition, debriefing texts shown in Table 3 in the main text. For now, we have only one human coder, but intend to expand this group to at least 5 in a future draft.
    c. Show human coders the texts in random order, **with** the algorithm sentiment scores attached, and ask them to approve or disapprove (binary) of each text's score per our sentiment definition. If they were unsure, they were asked to still provide their best guess.
    d. Consolidate human-coded texts by taking the raw average of approval ratings across coders (0 is disapprove, 1 is approve)
    e. Calculate approval rates overall and by each grouping of algorithm-coded sentiment score.

VIIIe. Assessing the Confidence of Random Forest Sentiment Scores

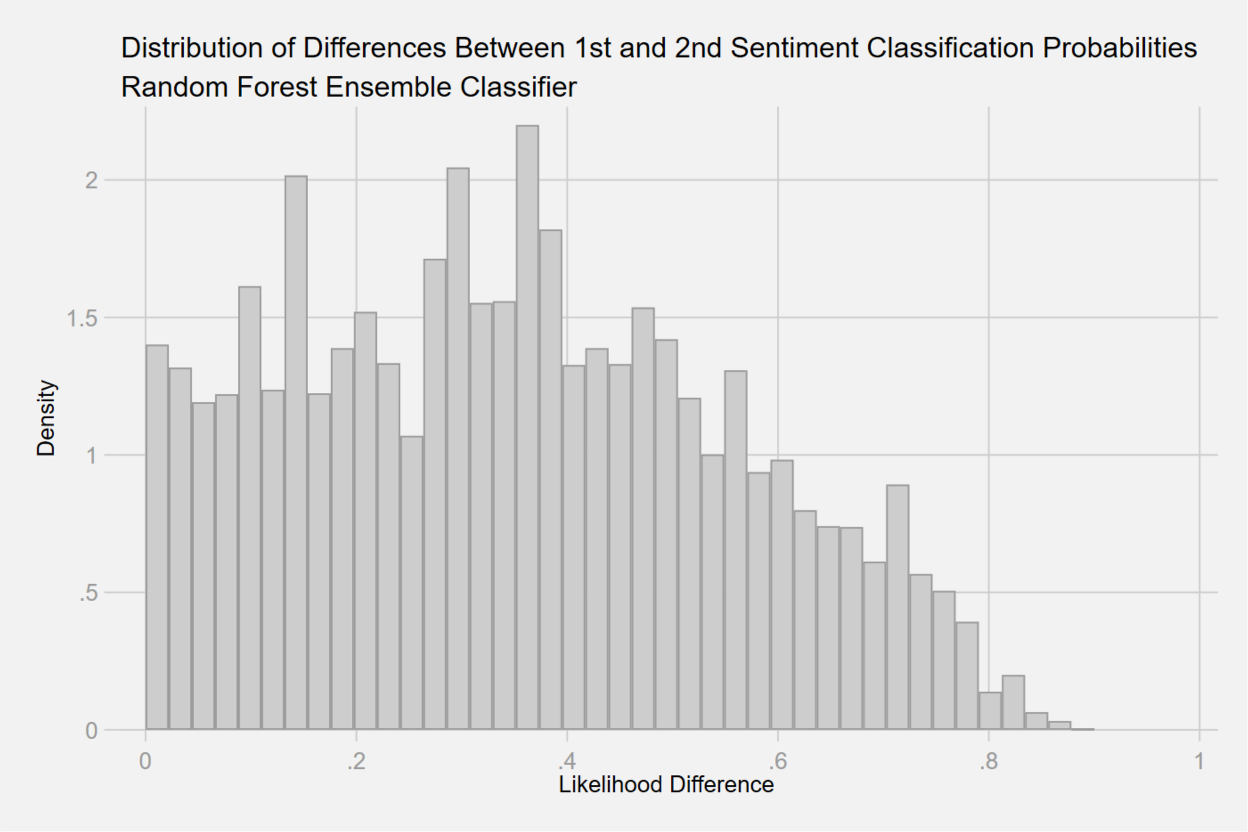
As mentioned in the narrative above, an important diagnostic to evaluate here is how "certain" the algorithm is when making a decision. Our ensemble outputs how likely it thinks *each* possible sentiment score is for a given sentence. Figure E1 shows the associated probability of each of the*final* classifications the algorithm has provided; roughly, how confident it was in each individual classification. We see that the mean sits around 60%, meaning the algorithm is quite certain. We'd be worried if the mean were closer to 20-30% (given that complete uncertainty would produce a 20% probability across each of the five possible classifications).

Figure E1. Distribution of Probabilities for Assigned Sentiment Scores



We can also compare the likelihoods of the algorithm's first- and second-choice sentiment classification to see how close the two are. The closer they are, the harder a time the algorithm is having while picking between its best options. Figure E2 plots the *difference* between the probability of the algorithm's first-choice and the second-choice predictions across N2FL texts. The mean and median difference is approximately 34 percentage points, indicating that for the majority of texts, the algorithm is quite certain that its final classification is by far the best one. However, there is still a substantial volume of texts where the difference is negligible, and we may consider handling those predictions differently than the others in future drafts for this reason.

Figure E2. Distribution of Differences Between 1st and 2nd Sentiment Classification Probabilities



VIIIf. Topic Modeling Introduction and Methodology

Topic modeling is another common task in NLP in which analysts use an algorithm to "learn" what discrete topics of discussion exist across a series of documents and then measure how prevalent each topic is within each document. Ultimately, we are interested in whether there exists variation in the prevalence of these topics across students. Such variation would potentially reflect substantively different advising interaction content and arguably different "treatments" as a result.

To accomplish this, the standard topic modeling algorithm takes a large body of documents (in our case, text conversations) and attempts to analyze those documents for groupings of words that frequently co-occur together in the same documents (Blei, 2003). Words that frequently co-occur in this framework are thought to belong to the same abstract topic of discussion, and word groups identified by the algorithm in this way can then be interpreted by analysts for its substantive meaning. For example, the words "financial," "aid," and "loans" may all occur together at high frequency across various student-advisor conversations; the topic modeling algorithm would group these words together under one "topic," which could be interpreted by an analyst as the abstract topic of "financial aid information." The algorithm identifies several such topics based on the provided text, which allows for uniquely context-sensitive and flexible output compared to similar content analysis methods that use pre-defined word groups (e.g. the Linguistic Inquiry and Word Count method per Tausczik & Pennebaker, 2010). Once topics are identified, the algorithm can determine the prevalence of each topic in each conversation based on the combination of words in that conversation. For example, the algorithm can tell us how many words in a given conversation are spent discussing "financial aid information" as a topic, versus another topic that it discovers such as "student sports involvement."

For the present study, we opt to utilize the Structural Topic Model implementation proposed by Roberts et al. (2019). This methodology offers a refinement on the traditional topic modeling approach in a variety of ways, but most importantly for our purposes, it allows us to specify topical prevalence covariates as part of the topic modeling process. Put simply, this feature lets the topic model discover associations between the provided covariates and the prevalence of each topic to inform model fit. For example, if students who are older systematically discuss financial aid more often than students who are younger, the structural topic model can pick up on this and form more accurate expectations of how prevalent that topic is for all older students' conversations.[14] For our model, we use the following student baseline covariates for our topical prevalence covariates: institution, state, sex, race/ethnicity, age (over/under sample median of 23 years old), BA transfer intention, and prior transfer status.

There are three main challenges with regards to robustness and usability in topic modeling output. First, topic modeling is highly sensitive to the structure of the raw data and produces poor results when the length of the text documents are **(a)** too short (a good rule of thumb in practice is to use documents about the length of a paragraph), and **(b)** when the text documents simultaneously cover too many topics. In our case, single messages alone are likely too short a document size, while compiling all messages sent between a student and their advisor together likely covers too many distinct topics at once. As such, we chose to structure the data at the *conversation* level, defined as any messages sent by either the student or their advisor after a scheduled message, but before the next scheduled message. Given that we have good reason to expect the group of messages following a scheduled message would be related in content (e.g., a scheduled message about the FAFSA is likely followed by a student-advisor conversation about financial aid), and that grouping messages in this way would increase the length of each text document, condensing messages to the conversation level nicely addresses both of these concerns. We moreover restricted our training set to only personalized messages to prevent the content of scheduled messages from having undue influence on topic formation.[15] Thus, our topic modeling training set includes only conversations from students who responded to at least one scheduled message, for a total of 16,828 unique conversations with an average length of 19 "keywords."

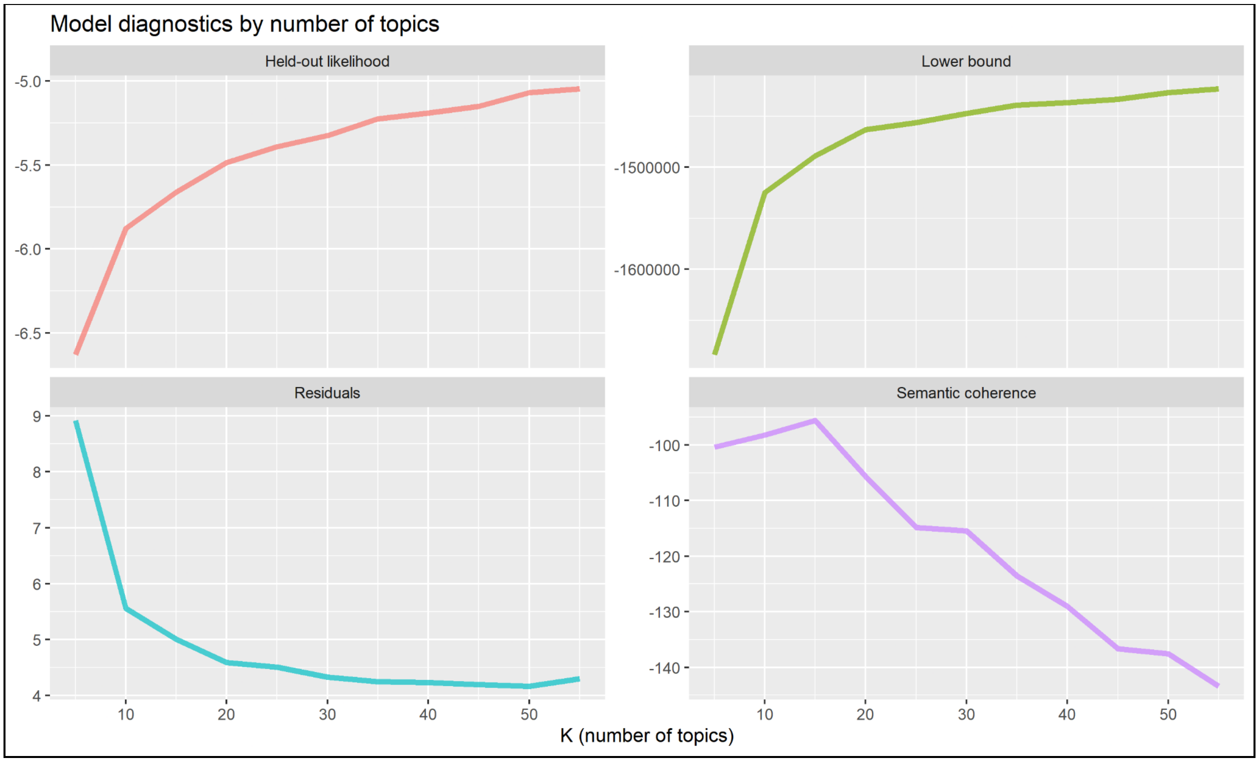
Second, topic modeling results are highly sensitive to the number of topics the algorithm is asked by the analyst to identify, known as the K parameter. Standard practice in the field is to run the topic model several times while arbitrarily changing K across a wide range of values, and then selecting the final model's K parameter based on a variety of model fit metrics (Roberts et al., 2019).

[14] In a more technical sense, the entirety of the topic modeling process is rooted in Bayesian frameworks, such that a word is really assigned a *probability* of being "about" each topic, and each conversation is assigned a *probability* of discussing each topic. Topical prevalence covariates allows the model to adjust a conversation's *prior* distribution of discussing each topic based on those covariates, which then shapes its *posterior* distribution of discussing each topic given the words within it and that prior distribution.

[15] To pre-process the text, we also **(1)** removed any stopwords, numbers, URLs, and other non-English language, **(2)** replaced proper nouns with stand-in tokens (e.g. "tokensystemname" instead of "John Jay") to increase language uniformity across contexts, and **(3)** manually spell-checked and aligned word/verb forms for the 3000 most common words across the dataset. We moreover include both unigrams (single words) and bigrams (common two word pairs) in our model, and exclude any tokens that occur in fewer than 10 (for unigrams) or 20 (for bigrams) documents to reduce the sparsity of the model.

We test every multiple of 5 up to 55 for our K parameter and find that K=30 produces the greatest balance of general model fit (held-out likelihood and model residuals) against an algorithmically-derived measure of word coherence within topics (semantic coherence, per Mimno et al., 2011). The results of our specification test are visualized in Figure F1 below.

Figure F1. Common Model Fit Metrics Across K Parameter Specifications



Lastly, the topic interpretation process can be highly subjective given that it is up to analysts to determine the substantive meaning (if any) in the topic groupings. To address this directly, we set up a multi-stage, multi-coder process to interpret the topic groupings inspired by Penner et al., 2019. We began by providing three coders (each of the co-authors) with a list of the words that most distinguished each of the 30 topics.[16] Each coder was asked to identify a unifying idea or concept for each topic to the best of their ability. Once each coder completed this process independently, we reviewed any disagreements and discussed how to harmonize these interpretations collaboratively. We found that all three coders had **perfect or near-perfect agreement on 26 out of 30 topic interpretations**, with the remaining 4 showing only minor disagreement. Complete harmonization tables and process documents are available upon request. Table F1 displays example topics with

[16] Coders received both the highest probability and highest frequency/exclusivity (FREX) words for each topic. The highest probability words are those that the algorithm thinks are most likely to belong to a given topic when they appear (i.e. the strongest indicators for a topic being discussed). The highest FREX words are words that are both highly frequent, *and* highly exclusive to that particular topic, in that they don't tend to appear in other topic groupings (Airoldi & Bischof, 2016). Balancing exclusivity with frequency is important to focus on words that matter in the documents; terms with high exclusivity but low frequency tend to have very little impact on the algorithm overall, and tend to be highly noisy.

perfect agreement, near perfect agreement, and minor disagreement among coders. Moreover, the 4 minor disagreements were easily resolved with brief clarification of the task and our written interpretations, and were ultimately inconsequential given our next step of combining related topics into larger "supertopics."

Table F1. Sample Topic Interpretations and Word Groups

Distinguishing Words	Coder 1	Coder 2	Coder 3	Coder Agreement
graduation, apply, apply graduation, express, tokensystemname express, application, link, walk, ceremony, cunyfirst	applying to graduation	Graduation application	apply for graduation	Perfect
office, hour, late, friday, monday, response, stop, visit, late response, c107	finding a time to visit a campus office	office hours	[office] contact logistics	Near-Perfect
campus, service, counselor, job, support, mind, ahead, provide, care, set	campus resources to support students	work	Counseling services	Minor Disagreement

Following the interpretation of individual topics, each coder then individually combined topics into broader supertopics to reduce the dimensionality of our topic model output and improve our ability to relate our output to substantively relevant advising practices (e.g. share of conversations focusing on financial aid versus course registration detail).[17] In a similar manner to our initial topic interpretation process, we then harmonized the supertopic groupings across each of the coders' proposed schemes. Our harmonized supertopics are displayed in Table F2 below. Note that four of the 30 topics were not ultimately grouped into an actual supertopic for analysis due to their lack of substantively relevant meaning (e.g. pleasantries like "hey, i'll, glad, yeah, awesome, haha, alright" or more basic communication logistics like "email, student id, check, text, stop" etc.).

Table F2. Complete Supertopic Groupings and Sample Words

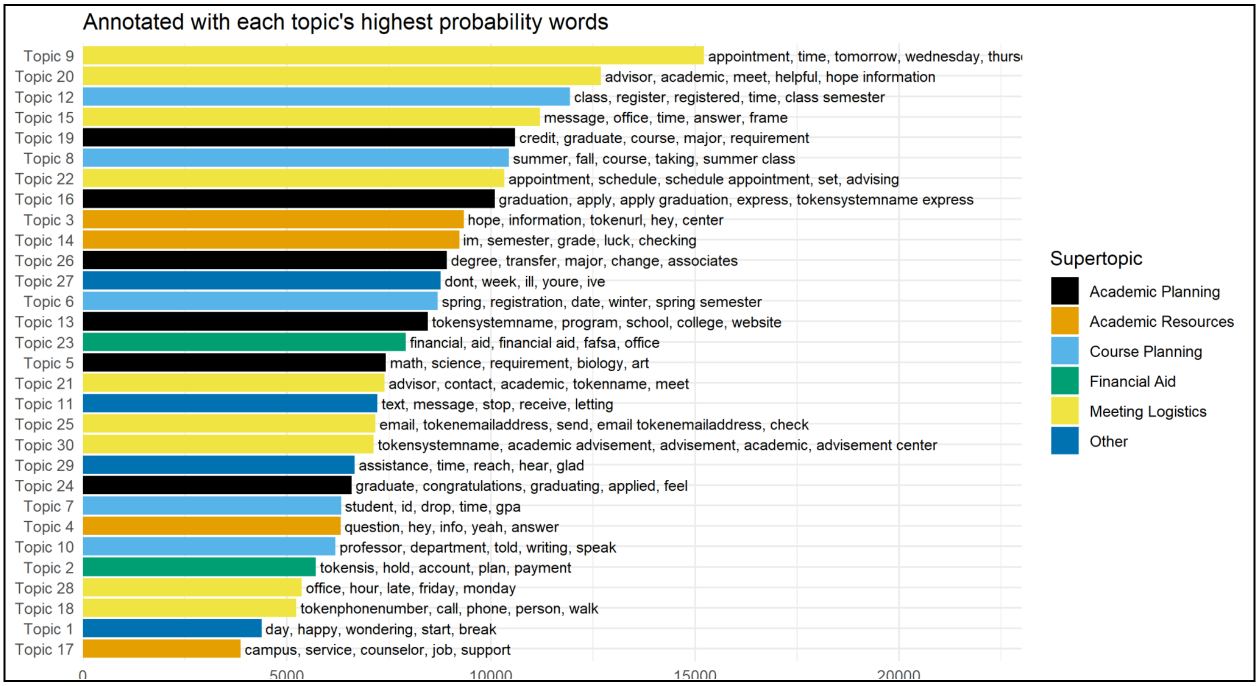
Academic Planning	math, science, requirement, biology, art, spanish, registrar, language, college, mat
	tokensystemname, program, school, college, website, tokensystemname tokensystemname, nursing, online, application, tokensystemname website
	graduation, apply, apply graduation, express, tokensystemname express, application, link, walk, ceremony, cunyfirst
	credit, graduate, course, major, requirement, internship, psychology, minor, elective, missing
	graduate, congratulations, graduating, applied, feel, free, ready, december, feel free, graduated

[17] Mathematically speaking, we are considering the probability that a given document discusses each supertopic a shorthand for "discussing topic A, OR discussing topic B, OR discussing topic C," etc. Thus, the probabilities that each document discusses each topic are summed to the probabilities that each document discusses each supertopic instead.

	degree, transfer, major, change, associates, plan, audit, transcript, bachelors, finish
Academic Supports	hope, information, tokenurl, hey, center, tutoring, located, helpful, office, visit
	question, hey, info, yeah, answer, reaching, nice, assist, specific, study
	im, semester, grade, luck, checking, enrolled, final, exam, planning, lol
	campus, service, counselor, job, support, mind, ahead, provide, care, set
Meeting Logistics	appointment, time, tomorrow, wednesday, thursday, tuesday, monday, meet, availability, day
	message, office, time, answer, frame, time frame, message time, answer message, frame patience, patience
	tokenphonenumber, call, phone, person, walk, call tokenphonenumber, monday, hour, plan, discuss
	advisor, academic, meet, helpful, hope information, information helpful, appointment, academic advisor, meet academic, hope
	advisor, contact, academic, tokenname, meet, advising, academic advisor, track, meet advisor, reach
	appointment, schedule, schedule appointment, set, advising, advisor, tokenurl, link, met, meet
	email, tokenemailaddress, send, email tokenemailaddress, check, received, information, connect, forward, contact
	office, hour, late, friday, monday, response, stop, visit, late response, c107
	tokensystemname, academic advisement, advisement, academic, advisement center, center, line, reach, call, tokenphonenumber
Course Planning	spring, registration, date, winter, spring semester, november, enrollment, session, register, semester
	student, id, drop, time, gpa, student email, check, access, withdraw, student id
	summer, fall, course, taking, summer class, online, fall semester, summer course, plan, im taking
	professor, department, told, writing, speak, permission, request, alright, issue, morning
	class, register, registered, time, class semester, add, pin, class im, register class, left
Financial Aid	tokensis, hold, account, plan, payment, pay, bursar, log, check, tokenurl
	financial, aid, financial aid, fafsa, office, aid office, scholarship, loan, tuition, pay
	day, happy, wondering, start, break, due, hope, yesterday, deadline, january
	text, message, stop, receive, letting, update, wrong, list, text message, send
	dont, week, ill, youre, ive, havent, sounds, awesome, didnt, fine
	assistance, time, reach, hear, glad, taking, text, respond, hesitate, taking time

Figure F2 below displays the prevalence of each underlying topic, and its corresponding supertopic, in terms of word frequency within the training dataset (personalized messages between students and advisors, collapsed to the conversation level). Note that the process of deriving the number of words in each conversation that come from each topic is probabilistic in nature. That is, because a single word can belong to multiple topics at once (e.g. "deadline" might appear in financial aid and in course enrollment discussions) at varying probabilities (perhaps it is more common in financial aid than course enrollment), the algorithm will use these probabilities to assign it to a topic each time the word appears. The algorithm runs many simulations given these parameters and the input text, and the output topic assignments are the modal value from the distribution of those words to topics across simulations.

Figure F2. Topics and Supertopics by Frequency of Word Occurrences in Topic Model Training Data



To summarize, our topic modeling process allows us to estimate, for each student, the share of their conversation focused on each supertopic of conversation:**(a)** financial aid, **(b)** advising meeting scheduling logistics, **(c)** course registration and enrollment, **(d)** broader academic planning, and **(e)** academic resources. While there is some overlap and close relationship between these supertopics in concept (e.g. some advising meetings are likely set up to discuss financial aid, or course registration, etc.), we argue that these present clearly delineated characterizations about the content of the text messages themselves and allow us to credibly characterize trends and variation in texting patterns across students as a result.