READING BETWEEN THE LINES: A TEXT ANALYTICS EXPLORATION OF SOCIAL EMOTIONAL LEARNING

A Thesis
by
ADAM SCOTT BLACKBURN

Submitted to the School of Graduate Studies
at Appalachian State University
in partial fulfillment of the requirements for the degree of
MASTER OF ARTS

May 2022
Department of Psychology

READING BETWEEN THE LINES: A TEXT ANALYTICS EXPLORATION OF SOCIAL EMOTIONAL LEARNING

A Thesis
by
ADAM SCOTT BLACKBURN
May 2022

APPROVED BY:

Shawn M. Bergman, Ph.D.
Chairperson, Thesis Committee

Rose Mary Webb, Ph.D.
Member, Thesis Committee

Yalcin Acikgoz, Ph.D.
Member, Thesis Committee

Rose Mary Webb, Ph.D.
Chairperson, Department of Psychology

Marie Hoepfl, Ed.D.
Interim Dean, Cratis D. Williams School of Graduate Studies

Abstract

READING BETWEEN THE LINES: A TEXT ANALYTICS EXPLORATION OF SOCIAL EMOTIONAL LEARNING

Adam Scott Blackburn
B.A., University of North Carolina Wilmington
M.S., University of Florida
M.A., Appalachian State University

Chairperson: Dr. Shawn Bergman

Research has shown that social emotional learning (SEL) is being increasingly valued by schools due to its associations with academic achievement and student engagement. Unfortunately, the current state of SEL, with its lack of time or resources allocated by practitioners and prevalence of self-report assessment, restricts scalability and does not meet the forecasted demand. Additionally, despite the importance of assessing SEL, difficulties in measuring it persist. The lack of consensus contributes to misinterpretation, over-generalization, and/or overlooking evidence to help with further developing strategies for measurement and evaluation. This paper will present that SEL assessment needs to be innovated to help address the limitations of self-report measurements and necessity for a scalable solution. Text analytics and natural language processing (NLP) serve as a flexible, low-lift assessment method that can analyze contextual and subgroup differences addressing these limitations of existing SEL assessment. Therefore, this study presents a text analytics and NLP evaluation of a proposed text-based SEL assessment of growth mindset assessing if analysis of text message conversations between agents and students can be used to assess a student's level of SEL. Conducting a review of the relevant literature, to the best of our knowledge this is the first study to assess a text-based SEL assessment using a text analytics and NLP approach. Ultimately, this study created five prediction models for growth mindset scales with predictive validities between .37 and .43.

Keywords: social and emotional learning; growth mindset; text analytics; natural language processing; word embeddings; BERT

Acknowledgments

My thesis committee chair, Dr. Shawn Bergman, has been instrumental in advancing my thesis timeline and assisting with continuous feedback. Chad Heinen and Justin Cyrus of Student Success Agency have also provided an immense amount of support by providing data, analysis suggestions, and access to their systems. Furthermore, Mr. Heinen helped pull and organize the data for analysis, and Mr. Cyrus helped find and train agents to participate in the study. Additionally, Dr. Scott Marshall and Lindsay Marshall have both been instrumental in learning to navigate large datasets and prepare the appropriate text analysis methodology. I would also like to thank Jennifer Wilson-Kearse, and all the members of the Vela Institute, who have assisted with the data and funding for this project. The time and effort of Dr. Rose Mary Webb and Dr. Yalcin Acikgoz has also been greatly appreciated. Natalie Walker's hours of proofreading my previous drafts has been very illuminating. All members have devoted a large amount of their time and resources to assist my goals in finishing the project, and for that I am grateful.

Dedication

To being waitlisted on my first application to Industrial Organizational Psychology and Human Resource Management; in the words of Douglas Adams, “I may not have gone where I intended to go, but I think I have ended up where I needed to be.” Also, to my family, I appreciate your support, sometimes unrecognized and under-rewarded, of my journey as a “career student” to help me determine and pursue my academic passions.

Table of Contents

List of Tables

List of Figures

Reading Between the Lines: A Text Analytics Exploration of Social Emotional Learning

An increasing trend in education is to focus on social and emotional learning (SEL) due to its associations with academic success and student engagement (Ashdown & Bernard, 2011; Denham & Brown, 2010; Eklund et al., 2018; Hart et al., 2020; Jones et al., 2014; McCormick et al., 2015; NJSBA, 2017; Yang et al., 2018). In fact, every state has free-standing preschool SEL standards, and 11 states have freestanding SEL standards at the K–12 level (Eklund et al., 2018). These standards enable the states to develop accountability systems for assessing student success by guiding programmatic SEL efforts in schools. Given this interest, numerous SEL frameworks and definitions have been proposed which can contribute to confusion in comparing assessments, interventions, and approaches. Researchers are now attempting to address this existing confusion by creating summative frameworks, emphasizing balancing research rigor with practitioner relevance, and proposing alternatives to the predominate self-report rating assessments (Ecological Approaches to Social Emotional Learning [EASEL] Lab, 2021b; Elias, 2019; Liu & Huang, 2017; McKown, 2019). As this paper will present, assessments need to be scalable due to the rising interest in SEL. Self-report rating assessments limit scalability (Elias, 2019); therefore, the field needs to focus on innovative SEL assessments.

It is important to assess SEL as it encompasses student skills that tests of academic achievement and ability do not encompass (West et al., 2020). Lawson et al. (2019) showcased that the core components across SEL programs were social skills, identifying others' feelings, identifying one's own feelings, and behavioral coping skills/flexibility. These interpersonal and intrapersonal skills predict a variety of academic and life outcomes, portraying student success inside and outside of the classroom (Almlund et al., 2011; Deming, 2017; Heckman et al., 2014). Furthermore, schools should support this additional method of assessing student success; for example, the federal Every Student Succeeds Act (ESSA) requires states' school accountability systems to include an additional indicator of school quality or student success beyond their academic scores (Eklund et al., 2018; West et al., 2020). SEL can serve as such an additional indicator. Thus, it is important to assess SEL as it can advise state-level and school district curriculum and policies.

Despite the importance of assessing SEL, difficulties in measuring it persist. These difficulties may persist due to SEL being a cross-disciplinary topic. Jones et al. (2019) describes a lack of consensus among researchers, practitioners, policymakers, and frameworks from different disciplines resulting in a jingle-jangle fallacy of referring to the same competency or skill by different names, or alternatively, the same name for two unrelated competencies or skills. This lack of consensus contributes to misinterpretation, over-generalization, and/or overlooking evidence to help with further developing strategies for measurement and evaluation. The Harvard Taxonomy Project is one attempt to address this dilemma and summarizes that SEL "has often been used as an umbrella term to represent a wide array of non-academic skills that individuals need in order to set goals, manage behavior, build relationships, and process and remember information" (EASEL Lab, 2021b, para. 2). Currently their database highlights 40 different SEL frameworks permitting researchers the ability to see similarities and differences in terminology (EASEL Lab, 2021a). They have classified all of these frameworks across six domains: cognitive, emotion, social, values, perspectives, and identity. Within these domains, subdomains are provided helping to conceptually define these domains (detailed in Table 1), and within each subdomain there are competencies providing further conceptualization and operationalization clarification (EASEL Lab, 2021a). Despite this progress to begin creating a concrete framework, jingle-jangle fallacies within subdomains and competencies of SEL still exist and practitioners need guidance on which assessments to utilize.

With terminology and measurement difficulties existing throughout SEL, which are partly being addressed through efforts like the Harvard Taxonomy Project, researchers should also look to develop an innovative way to assess SEL beyond creating yet another self-report framework which may perpetuate these difficulties. The field needs to address the difficulty in feasibly scaling up high-quality SEL given that more school districts are wanting to assess SEL. Unfortunately, there are not enough SEL experts to satisfy this demand (Elias, 2019). While one does not need to be an SEL expert to assess SEL, experts assist with navigating the terminology difficulties and are vital for helping evaluate and validate assessment approaches for existing and novel interventions based on a districts' research goals.

SEL assessment must be feasible for all existing and novel interventions, delivering value to practitioners while satisfying research rigor (McKown, 2019). Current traditional SEL programs primarily consist of scripted and sequenced curricula designed to be implemented during the school day, and deviations from this script can be seen as threats to program fidelity (Bailey et al., 2019). With this current state of SEL, schools will need to invest in continuous staff training and professional development (Elias, 2019; Jones & Bouffard, 2012). Additionally, these approaches fail to leverage the expertise of teachers who know the students in their classroom best and provide autonomy in utilizing their relationships, observations, and decisions towards delivering effective and timely SEL support to students. Therefore, some are calling for SEL interventions that permit more flexibility beyond a set curriculum, letting practitioners determine which students need what SEL interventions at what times (Bailey et al., 2019). Current assessment methods place more emphasis on curricula and training people to follow the curricula rather than providing tailored solutions to students. With the increased demand for SEL, current experts are only able to satisfy the predominate curricula emphasis perpetuating a gap in assessment of the flexible intervention approach (Elias, 2019).

Suggestions for modifying SEL assessment from its current state include focusing on flexible, low-lift strategies and practices, not just curricula, and ensuring teachers are responsive to students' specific needs and experiences (Bailey et al., 2019). A low-lift strategy refers to designing an approach that is flexible enough to be implemented at any time of day and in any context (e.g., class, lunch, hallways, recess). Using this style of approach permits one to address challenges or opportunities as they arise rather than waiting for the relevant content to come up in the curricula (Bailey et al., 2019). Additionally, a low-lift approach may aid in addressing the limitation of insufficient time to devote to SEL during the day through abridged or skipped curricula lessons (Jones & Bouffard, 2012). Moreover, teacher responsiveness helps plan SEL instruction which reflects students' lived experiences in their communities, at home, and at school (Bailey et al., 2019). Tailoring the instruction can assist with buy-in and participation from students (Elias, 2019). Thus, assessment needs to be able to discern which students need a targeted intervention while also acknowledging environmental and contextual effects on SEL competencies.

When evaluating new assessment approaches, it is important to note SEL is a broad area; therefore, a few SEL areas should be selected and analyzed first (Liu & Huang, 2017). Further, it is important not to overgeneralize the feasible uses of assessment approaches (McKown, 2019). Calling upon these recommendations, a well-researched area of SEL is growth mindset, the belief one can develop their most basic abilities through dedication and hard work (Dweck, 2017). This is an appropriate first area of SEL to evaluate new assessment approaches within because its contextual effects have been documented (Claro et al., 2016; West et al., 2020). Thus, there should not be significant demographic differences in expression of growth mindset permitting more attention to be on environmental factors in low-lift assessment approaches.

Reviewing the state of SEL research, McKown (2019) highlighted five challenges and opportunities in applied assessment of student SEL. First, researchers need to integrate developer and practitioner priorities focusing on combining psychometric rigor with practical relevance. Second, goal clarity is needed for the interpretation and use of SEL assessment data. Third, there is a challenge to create conditions for high-impact SEL assessment; "data can be useful only if users have an opportunity to review, discuss, and interpret their meaning to decide what to do based on what they learned" (McKown, 2019, p. 213). This challenge is where teacher input would greatly be appreciated; teachers can leverage their insights to ensure their SEL interventions are responsive to their students' needs and experiences. Fourth, the field needs to coordinate standards, programs, assessment, and professional learning. Lastly, there needs to be more investment in unwavering assessments in varied cultural contexts. To address these challenges, suggestions include involving the intended end-user in the assessment development and utilizing text analytics to go beyond Likert question surveys (Elias, 2019; McKown, 2019). Text analytics and natural language processing could serve as a flexible, low-lift assessment strategy and such a solution for these challenges.

Text analytics is the automated process of translating text into data which can be analyzed for patterns, trends and insights (Lexalytics, 2021; Pustejovsky & Stubbs, 2012). Natural language processing (NLP) is designed to facilitate human interaction with machines and devices, giving computers the ability to process human language and understand its meaning, including intent and sentiment (Pustejovsky & Stubbs, 2012). One such method of NLP is word embeddings which permits machines to understand the representation and meaning of the word based on its context in the sentence and entire text document (Brownlee, 2017; Devlin et al., 2019). The ability to determine intent and sentiment, while also understanding the context for when a word was used, will further help identify patterns and trends, or lack thereof, in the display of SEL competencies. Together, text analytics, natural language processing, and word embeddings are an analytical method that involves creating an algorithm to help design a model that can discern differences between groups.

This approach helps address the issues presented by McKown (2019) due to its ability to let researchers and practitioners collaborate on determining what to include within the algorithm as well as providing guidance on how to interpret its results. This approach also enables one to efficiently retrieve information from large datasets (Al-Maitah, 2018). This technique, analyzing already collected data and readily accepting new data, permits a continual opportunity for users to review, discuss, and interpret across time (Liu & Huang, 2017). Furthermore, as text analytics is a digital, analytical assessment method and technology is increasingly being integrated into classrooms, scalability is easier. With enhanced scalability, one will be able to assess competencies across varied environmental and cultural contexts. Additional variables can be added to the algorithm to help begin to address the differences caused by environmental or cultural contexts and their intersectionality (Elias, 2019). For example, Ferreira et al. (2018) analyzed student discussion messages in different sections of a master's level software engineering course to determine the relationship between different course topics and students' cognitive presence; they then added "instructional scaffolding" (such as a role-assignment intervention) to their algorithm to assess how cognitive presence changed in the presence or lack of "instructional scaffolding" in their classroom environment.

Acknowledging the need for SEL innovation and methods to address the existing challenges in the field, the purpose of this study is to evaluate if text analysis and NLP of text conversations with students can be used to assess a student's level of SEL. Previous research has focused strongly on assessing and forming various self-report SEL measures and frameworks common in traditional SEL interventions. This exploratory study is used to help design an assessment approach that will map across SEL subdomains and competencies and be feasible to assess traditional and flexible, low-lift interventions. This approach will lead to descriptive results helping discern what differences exist, begin to identify and label differential factors, and begin to understand where they occur. Criterion validity will be established with mentor rankings. Differences in SEL scores will be utilized to identify student characteristics that relate to overall SEL.

Review of Relevant Literature

Social and Emotional Learning

SEL comprises the processes through which children and adults learn to regulate and understand their emotions, strengthen the competencies that are the foundation for human interaction, establish goals, process and remember information, and make responsible decisions (EASEL Lab, 2021b; Eklund et al., 2018; Elias, 2019). These capabilities encompass numerous skills. Practitioners and researchers have emphasized SEL due to its association with student academic success, student engagement, and teaching efficacy. For example, Ashdown and Bernard (2011) analyzed preparatory and grade one students attending a low socio-economic status Catholic school in Australia and asked teachers to rate students' social-emotional well-being before and after a 10-week SEL classroom program. Their results indicated that students who completed the SEL classroom program displayed growth in SEL competence, reduction in counterproductive behaviors (e.g., hyperactivity, internalizing), and an increase in reading achievement for lower achieving grade one students. Analyzing a US sample of kindergartners and first graders, McCormick et al. (2015) found first-grade math achievement was mediated through improvements in classroom emotional support and organization while first-grade reading achievement was mediated through improvements in class organization. No mediation occurred for kindergartners between SEL and academic performance. Acknowledging SEL skills can vary within persons and across environments, Denham and Brown (2010) reviewed the literature of links between SEL and preschool and grade school academic performance. They summarized that a student's self-awareness may lay the groundwork for future academic accomplishments as it may provide them with the needed confidence to take more risks in the classroom in both participation and attempting challenging academic problems. Furthermore, a meta-analysis by Durlak et al. (2011) detailed students receiving SEL programs demonstrated gains in academic achievement (an 11-percentile gain, Hedges' $g = .27$) and social-emotional skills; therefore, it should be an essential aspect of pre-K-12 education. Assessing the impact of a universal SEL program on future Pennsylvania state test performance for third-fifth graders, Hart et al. (2020) found small effect sizes on reading and math scores (Hedges' $g < .2$) and saw the magnitude of effect appeared to be slightly larger for fifth grade proficiency status. Hence, it is apparent SEL has been linked to academic achievement.

Beyond academic achievement, SEL has also been linked to improvement in classroom management and student engagement, the student's connection or involvement with the people, activities, goals, and values of schooling (Yang et al., 2018). It is important to improve classroom management, maintaining attention of all students by redirecting negative or distracting behavior and pulsing the room to optimize student motivation and engagement. SEL can assist with classroom management by supporting students to manage themselves throughout daily classroom instruction. Those with higher SEL are found to be less disruptive and more attentive to classroom instruction (Jones et al., 2014). Additionally, training teachers with an SEL perspective assists with viewing disruptive behavior as a learning opportunity rather than an annoyance/hinderance. Assessing a created SEL program to facilitate classroom management, Jones et al. (2014) found more positive, emotionally supportive, and well-managed classroom environments as well as increased students' attention skills and reduced impulsive behaviors.

Yang et al. (2018) assessed school-wide SEL learning approach's impact on elementary, middle, and high school student engagement by analyzing teacher–student relationships, student–student relationships, and teaching of social and emotional competencies. Results indicated all three factors support cognitive-behavioral and emotional engagement, with teacher-student relationships being the strongest. Focusing solely on the teaching of social and emotional competencies, at both the student and school levels of analysis, increased perceptions of learning SEL were positively and significantly associated with cognitive-behavioral and emotional engagement. However, there was a weaker, but still positive effect in high schools than in middle or elementary schools (Yang et al., 2018).

While these findings with academic achievement, student engagement, and teacher efficacy have been discovered, various measurement methods of SEL have contributed to slightly different results (Collie et al., 2012; Hart et al., 2020; McCormick et al., 2015; West et al., 2020). Many SEL studies rely on small convenience samples of students in narrowly tailored settings causing generalizability concerns; additionally, the utilized SEL constructs and measures make it difficult to compare results across studies (West et al., 2020). These different results may also be due in-part to some SEL frameworks having a one-size-fits-all approach while others emphasize approaches that are developmental, flexible, and responsive to local needs (Bailey et al., 2019).

A one-size-fits-all approach regards having a standardized curriculum for all students/schools that is predetermined and cannot be tailored to individual classroom needs (Lawson et al., 2019). For example, the SSIS-CIP Early Elementary Version (Elliott & Gresham, 2007) is a standardized approach that requires about 10-12 hours of instructional time and comprises 10 core units that teachers use to teach students skills identified as critical to success in the classroom (e.g., following directions, asking for help, staying calm with others; Elliott & Gresham, 2007; Hart et al., 2020). Within a standardized curriculum, an SEL component is not assessed until its “core unit” which may leave needs of individual students unanswered. Further, the need of repetitive assessment after each unit also requires a lot of time and resources that are not always able to be adequately provided (Lawson et al., 2019). By being context-restricted and heavily resource dependent, this approach is not conducive to flexibility, a key requirement of SEL measurement scalability.

Low-lift strategies are flexible and permit tailoring SEL to individual classroom needs (Bailey et al., 2019). For example, Denham and Brown’s (2010) SEL model is comprised of self-awareness, self-management, social awareness, responsible decision making, and relationship skill assessments. It also emphasizes environmental conditions heavily impact SEL skill development. Their model permits variability to address individual classroom needs such as classroom management and peer effects (Denham & Brown, 2010; Jones et al., 2014). Low-lift SEL strategies may help address limitations of one-size-fits-all SEL assessment strategies like generalizability, difficulties being compared to other SEL assessment frameworks, and inability to attune to individual classroom needs leading to practitioners needing guidance on which framework to use for their intended purpose (Jones et al., 2019; Lawson et al., 2019; West et al., 2020). Thus, a low-lift assessment approach, like text analytics and NLP, begins to address these limitations as it is not restricted to classroom instruction; it can occur in any context, fostering a schoolwide commitment to assessing and supporting students’ SEL (Bailey et al., 2019). Conversely, researchers need to closely monitor reliability and validity of these approaches – too much flexibility in place of standardization could result in reductions to academic rigor, impacting practical relevance (Lawson et al., 2019; McKown, 2019). SEL experts can assist in monitoring and balancing standardization with flexibility. Furthermore, text analytics and NLP's potential integration within classroom technology can help reduce required resources and assess student differences, identifying why individual classrooms have different needs. This integration can help supplement and/or enhance self-report data from students.

The SEL literature also has a heavy reliance upon self-report assessments which have multiple limitations. As McKown (2019) describes, self-report assessments are vulnerable to a social desirability response bias and rely upon self-awareness of the individuals rating themselves which may not always be present. Social desirability bias regards answering in the way one thinks the evaluator wants rather than answering truthfully. Students may engage in this bias due to lack of buy-in to a one-size-fits-all curricula or lack of rapport with their teacher and not wanting to have follow-up conversations on these topics (Bailey et al., 2019; Elias, 2019). Additionally, there may not be enough time in the school day, or time in the school's curricula to allow students the time to complete a questionnaire contributing to low sustainability of SEL interventions (Bailey et al., 2019; Elias, 2019; Hart et al., 2020). Further, these questionnaires require adequate response rates to provide accurate, summative information and have to be done repetitively to track changes in students' competencies across time. Providing an assessment and solution, such as a low-lift, text analytics and NLP approach, which can be implemented in any context is not limited by time restrictions. Its flexibility for assessment by any teacher or school staff member throughout environmental contexts aids in schoolwide consistency in promoting and evaluating SEL, and its ability to be engrained in technology can assist with continuous, passive collection of data rather than requiring time for yet another assessment.

Acknowledging these limitations of self-report measurements, some assessments have transitioned to teacher or mentor ratings in-addition to, or in-place of, self-report ratings (Ashdown & Bernard, 2011; Jones & Bouffard, 2012). However, teachers or mentors may not have a fully representative view of the student's SEL or have the time/resources to appropriately rate all of their students (Ashdown & Bernard, 2011; Bailey et al., 2019; Lawson et al., 2019). There is also the challenge of phrasing questions to younger participants due to difficulties of reading and cognitive demands (McKown, 2019). This difference in ability and phrasing slows generalized assessment of students' SEL as researchers need to ensure their rephrased questions still appropriately map out to their assessed SEL factors. These limitations highlight another need to move away from self-report and other-report measurement to an assessment approach that is not a time or resource burden on staff and students while also permitting flexible assessment across contexts.

To maximize SEL's associated benefits with student academic success, student engagement, and teaching efficacy, proper assessment must occur (Durlak et al., 2011; Yang et al., 2018). Current SEL assessment is predominately characterized as being reliant upon small convenience samples of students in narrowly tailored settings causing generalizability concerns; being standardized and not-tailorable to individual needs/differences; using self-report assessments vulnerable to social desirability bias; and facing time- and resource-restriction concerns (Ashdown & Bernard, 2011; Bailey et al., 2019; Lawson et al., 2019; McKown, 2019; West et al., 2020). These existing limitations indicate a call for innovation in SEL assessment. Some have proposed low-lift strategies, which are flexible and permit tailoring SEL to individual classroom needs (Bailey et al., 2019) while others have proposed teacher or mentor ratings in-addition to, or in-place of, self-report ratings (Ashdown & Bernard, 2011; Jones & Bouffard, 2012). Per Liu and Huang's (2017) recommendation of focusing on a few SEL areas and analyzing them first before trying to be comprehensive of SEL entirely, this report will present a focus on growth mindset. Through its flexibility and ability to integrate with technology, a text analytics and NLP approach can assess current SEL interventions, incorporate the low-lift and teacher/mentor rating recommendations, address the time- and resource-restriction concerns, and answer the call for innovation and scalability in SEL assessment beginning within growth mindset.

Growth Mindset

As previously mentioned, there are numerous conceptualizations of SEL contributing to difficulties determining what the "assessment of SEL" can or should encompass and be designed. Acknowledging the difficulty comparing results across studies due to different utilized constructs and measures as well as the need to provide practitioners with guidance and clarity on SEL terminology (Jones et al., 2019; Lawson et al., 2019; West et al., 2020), the Harvard Taxonomy project has consolidated the multiple SEL frameworks and their assessed skills into six domains: cognitive, emotion, social, values, perspectives, and identity (EASEL Lab, 2021a). Within these domains, the Harvard Taxonomy project provides subdomains helping to conceptually define these domains (detailed in Table 1), and within each subdomain there are competencies providing further conceptualization and operationalization clarification (EASEL Lab, 2021a).

The Harvard Taxonomy is helping the field advanced towards a more concrete conceptualization of SEL. For example, this project can help address issues presented by McKown (2019) by assisting with coordinating standards, programs, assessment, and professional learning. As SEL becomes more conceptualized, practitioners will be able to align terminology and engage in benchmarking across different assessments, and as this challenge is further addressed, more time can be devoted to an assessment approach that is not a time or resource burden on staff and students while also permitting flexible assessment across contexts. In the current study, this Harvard taxonomy was used to conceptualize growth mindset as it has consolidated definitions and competencies across numerous frameworks. This conceptualization of growth mindset was then used to assess its impact on academic performance.

The Harvard Taxonomy Project's identity domain concerns perceptions of oneself and one's autonomy. Its subdomains include self-knowledge, purpose, self-efficacy/growth mindset, and self-esteem (see Table 2). Highlighted below, there are four competencies of growth mindset (EASEL Lab, 2021a):

- Believes that intellectual abilities and personality traits are qualities that can be developed and improved
- Expresses confidence in oneself and one's ability to improve or succeed
- Sees challenges as things that one can take on and overcome with time and effort
- Belief that one has a choice (agency)

Growth mindset fits within this domain as this competency encompasses the idea that through dedication and hard work, people believe they can develop their most basic abilities (Dweck, 2017). Of all the concepts within SEL, growth mindset was selected to be focused on because its contextual differences have been well-documented (Claro et al., 2016; West et al., 2020) as well as its association with academic success (Claro et al., 2016; Hochanadel & Finamore, 2015; Kim & Park, 2020), a valued outcome by practitioners often considered in the formation of SEL assessment programs (Eklund et al., 2018; Elias, 2019; Jones & Bouffard, 2012; Lawson et al., 2019). Growth mindset is important as it is associated with having grit (Hochanadel & Finamore, 2015) and developing new interests (O'Keefe et al., 2018). Researchers believe it can be cultivated and help one express their creativity. As contextual differences have been well-documented, it is important to provide a flexible, low-lift assessment strategy and provide teachers autonomy in utilizing their relationships, observations, and decisions towards providing effective and timely growth mindset support to students (Bailey et al., 2019). The ability to attune to contextual differences will also assist with scalability to other areas within SEL, areas that do not have as great of research as growth mindset. Hence improving assessment of growth mindset will be valuable to practitioners and serves as a focus area to integrate developer and practitioner priorities focusing on combining psychometric rigor with practical relevance per McKown's (2019) recommendations.

Alternatively, if a person is low on growth mindset, they may be considered to have a fixed mindset, where people believe their basic qualities, like their intelligence or talent, are innate and cannot be developed. They may also believe that talent alone, without effort, creates success (Dweck, 2017). Fixed mindset has been associated with negatively predicting academic achievement in elementary and middle school (Kim & Park, 2020) while also associated with low rates of retention in an online college environment (Hochanadel & Finamore, 2015). Assessing growth versus fixed mindset with an SEL approach helps schools understand their students' persistence in face of adversity (Hochanadel & Finamore, 2015; O'Keefe et al., 2018). A low-lift assessment approach provides the flexibility to assist with determining the various environmental factors a student has persisted through contributing to their display of a growth or fixed mindset.

Regarding contextual effects, there are minimal differences in growth mindset between boys and girls; patterns across racial/ethnic groups occur as early as middle school but narrow by more than half by grade 12 with increases seen for students of all races by grade 12; and growth mindset typically increases monotonically throughout elementary and middle school before leveling off in high school (West et al., 2020). Additionally, its positive relationship with achievement was found across all socioeconomic strata in a high school student sample from Chile (Claro et al., 2016). These findings highlight that growth mindset should be generalizable across subpopulations. If any significant subgroup differences exist, a text analytics and NLP approach can help discern why and where the differences exist.

With links to academic performance, persistence, and achievement, growth mindset is a valued outcome by practitioners because it helps answer who is successful and why. For example, growth mindset is associated with having grit which has helped predict who would get into and graduate from West Point, who would win the National Spelling Bee, and who would be a successful sales representative in a private corporation

(Hochanadel & Finamore, 2015). Naturally, all of these are different contexts where success can be uniquely conceptualized. Seeing this generalizability of growth mindset and grit, it is important to provide flexible assessment (Bailey et al., 2019). These findings highlight that growth mindset should be generalizable across contexts; if any significant context differences exist, a text analytics and NLP approach can help discern why and where the differences exist. Furthermore, given its value to practitioners, growth mindset provides a case study opportunity for researchers to integrate developer and practitioner priorities focusing on combining psychometric rigor with practical relevance.

Therefore, growth mindset is the focus of this study for a variety of reasons. First, it has a valuable association with academic performance, persistence, grit, and achievement (Claro et al., 2016; Hochanadel & Finamore, 2015; Kim & Park, 2020; West et al., 2020). Second, given its value to both students and teachers/mentors (Heggart, 2015), growth mindset is an ideal topic for researchers to integrate developer and practitioner priorities while focusing on combining psychometric rigor with practical relevance per McKown's (2019) recommendations. Third, as contextual differences in growth mindset have been well-documented (Claro et al., 2016; West et al., 2020) a flexible, low-lift assessment strategy should be used (Bailey et al., 2019). Thus, this study is exploring the feasibility of using text analytics and natural language processing (NLP) as a flexible, low-lift assessment of growth mindset that can discern differences between groups and account for contextual and environmental effects.

Text Analytics, Natural Language Processing, and Word Embeddings

Conceptual and Approach Overview

Text analytics and NLP is a method of analysis that helps address the limitations of self-report assessments (Elias, 2019; McKown, 2019). As mentioned previously, text analytics is the automated process of translating text into data which can be analyzed for patterns, trends and insights (Lexalytics, 2021; Pustejovsky & Stubbs, 2012). This automated process utilizes machine learning principles to derive meaning from unstructured text documents. That is, text analytics is the process of a computer transforming the raw text into usable information such as identifying who is talking, what they are talking about, and what they are saying about those topics (Mohler, 2020).

Text analytics' key goal of text quality evaluation helps one search information patterns as well as identify when matching semantic structures exist (Titenok, 2022). This approach assists with large data sets and when one needs to identify patterns across all of the qualitative data. For example, if a school district is trying to reconsider its elective offerings through an open-response survey, text analytics can list themes from all of the responses. Patterns in these themes can then emerge indicating what the most popular answers were, and then what were the most popular answers in specific subgroups such as by each individual school in the district or gender (Titenok, 2022).

Text analytics deals with the words in the text itself and employs a variety of tactics to quantify those words. Those resulting quantities are then used to better understand the concepts represented in the text and/or to predict a specific outcome. One such method text analytics utilizes is a bag-of-words approach that acts like a wordcount throughout the entire document. This approach removes the punctuation from the text and examines the occurrence of each word independently from the sentence in which it occurred. Thus, this approach, and most text analytics approaches, provides the frequency of words, not the context in which they occurred (Brownlee, 2019).

Text analytic techniques then takes these frequency counts to describe the natural text (e.g., the frequency of part of speech used, the pervasiveness of emotionally related words) and identify patterns automatically (e.g., if certain presence of certain themes were related to type of sentiment; Castillo, 2021; Linguamatics, 2021). For example, a Term Frequency-Inverse Document Frequency (TF-IDF) report can be created. Term frequency assesses the number of times a particular term appears within the document relative to the document's length. Inverse document frequency evaluates how common, or uncommon, a word is throughout the entire document or collection of documents and assigns a relative weight to the words; frequent words across all documents such as "a, to, and the" are given a lower weight, lower importance, to help uncommon terms (e.g., test, struggled, accomplishment) have a higher impact. Thus, following a bag-of-words and TF-IDF approach helps one quantify and evaluate the quality of the more important words in the text (Brownlee, 2019).

NLP is closely aligned with text analytics and while in many use cases it is difficult to differentiate the two, NLP is generally thought of as the next progression of text analytics where not only the frequency and relationships between words in a document are considered, but their relative context in the document is considered as well. Thus, NLP is a process that helps to further understand the underlying meaning of the information in the documents broken down by text analytics (Lexaltyics, 2020) by examining the text semantics and grammatical structures (Titenok, 2022). As a component of text analytics, NLP gives researchers the ability to process human language and understands its meaning, including intent and sentiment (Pustejovsky & Stubbs, 2012). For example, NLP is necessary to understand idiomatic expressions like “It was a piece of cake.” Text analytics evaluating each word independently would consider “cake” as a dessert rather than consider the entire phrase as saying the task was easy to do. NLP helps reveal the meaning of documents not only by evaluating word frequency but also by considering the grammatical structure of the documents (Titenok, 2022).

Assisting with the determination of intent and sentiment, NLP provides a variety of summary information including ontologies (creating a vocabulary of terms with constraints on its use) and semantic typing (identifying what a word or phrase-often found in an ontology-denotes; Linguamatics, 2021; Pustejovsky & Stubbs, 2012). Ontologies permit the ability to understand meaning even if the content is expressed in different ways or with a different context (Linguamatics, 2021). For example, when someone mentions “ASU,” ontologies can determine if they are referring to Appalachian State University or Arizona State University and permits NLP to understand App State also refers to Appalachian State University. Furthermore, ontologies can serve as the “vocabulary list” for the bag-of-words approach by providing the words the analysis is focusing on (Brownlee, 2019). These may serve as the rarer, prioritized words highlighted in the TF-IDF report. NLP’s ability to accurately summarize information and automatically identify patterns results in an algorithm that can apply these same decision-making principles to new sources of text data (Pustejovsky & Stubbs, 2012).

Once this vocabulary has been determined, semantic typing assists in determining what a word or phrase denotes. Examples of semantic typing include identifying persons, organizations, places, and times – helping understand who is talking, the content they are talking about, and the emotion/sentiment they express about that content (Mohler, 2020; Pustejovsky & Stubbs, 2012). For example, semantic typing helps NLP understand the difference in the use of "history" between "I have to go to history class" and "she and I have a bit of history together." Identifying these features assists in determining patterns and helps the computer learn if it needs to expand its categorization. For example, times can refer to days of the week or certain hours of the day. Through analyzing more data for patterns, NLP can learn if it is important to distinguish these types of time (Pustejovsky & Stubbs, 2012). The ability to create better categorizations assists the text analytics and NLP's algorithm in applying these same decision-making principles to new sources of text data.

A newer area of focus within NLP is word embeddings, which enhance the generalizability of results by acquiring a learned representation for text to provide similar representation to words with similar meetings (e.g., fight and brawl; Brownlee, 2017; Chen et al., 2013). In short, word embeddings assess the relationships between words and assess their similarity in meaning (Kjell et al., 2021). Word embeddings are based on a linguistic theory namely the "distributional hypothesis" (Harris, 1954), which broadly states that words that have similar context will have similar meanings. Word embedding algorithms determine word similarity by first assessing the probability that other words appear before and after a specific word a body of text (i.e., the extent to which a specific word is "embedded" in sentence). These probabilities represent the context in which a word tends to be used. The probabilities are then compared to other words determining which words in a language tend to be used in similar contexts (i.e., which words tend to have the same words coming before and after it). Words that are determined to be used in similar contexts (i.e., tend to have the same words preceding and following them) will have similar meaning.

The word embeddings approach allows for greater scalability by permitting text analytics for larger and smaller datasets as well as speeding up computation by facilitating the ability to assess the context of a word in-relation to the entire source of text (Wiedemann et al., 2019). Pre-trained models have the similarities and representations of a large corpus of words together which can be used on new data to relatively quickly convert text to numeric values that can then be used to predict outcomes. Thus, word embeddings can serve as a low-lift assessment strategy as it provides a continual opportunity for users to review, discuss, and interpret assessment results across time and contexts permitting enhanced scalability (Elias, 2019; Liu & Huang, 2017; McKown, 2019; Wiedemann et al., 2019).

Multiple text embedding models exist, such as Word2Vec (Mikolov et al., 2013), Global Vectors for Word Representation (GloVe; Pennington et al., 2014), and Bidirectional Encoder Representations from Transformers (BERT; Devlin et al., 2019). Word2Vec has served as the de facto standard for developing pre-trained word embeddings and permits better understanding of idiomatic phrases (e.g., “The New York Times” is an entity and should be considered as a whole; it should not be considered as the natural combination of the meanings of “new” “York” and “times;” Brownlee, 2017; Mikolov et al., 2013). A strength of Word2Vec is its requirement of less computational power compared to previous embedding approaches permitting higher-quality word embeddings and more dimensions of embeddings to be learned; to create the word embeddings, Word2Vec uses a continuous bag-of-words (CBOW) or continuous skip-gram model (Brownlee, 2017). Both use a word's local usage context (where it is nestled amongst its neighboring words) to learn about the word. The CBOW method predicts the current word based on its context while the continuous skip-gram model predicts the surrounding words based on the current word (Mikolov et al., 2013). This approach helps a machine learning algorithm create a more accurate predictive model for creating and understanding the embeddings due to its ability to predict each next word in the text document (Kulshrestha, 2019). Having more accurate embeddings will permit greater accuracy in follow-up analyses using the embeddings.

GloVE is an extension of Word2Vec that has improved its ability of capturing word meaning and demonstrating its meaning via tasks like calculating analogies (e.g., King – Man + Woman = Queen; Brownlee, 2017). Beyond just using the word's local usage context, GloVe analyzes the word's usage context based on the entire text corpus using matrix factorization techniques (Pennington et al., 2014). Unlike previous methodologies that considered the entire matrix of words, GloVe only assesses the nonzero elements in the matrix (i.e., examines only the words that appear in the text documents), which increases computational speed and accuracy of the resulting numeric word embeddings (Brownlee, 2017; Pennington et al., 2014). Both Word2Vec and GloVe are relatively equivalent at capturing semantic information, but GloVe tends to perform better at identifying word analogies and named entity recognition (i.e., determining if a word refers to a person's names, location names, company names, etc.; Brownlee, 2017; Pennington et al., 2014).

BERT is one of the latest extensions and advancements in the embeddings literature as it is a contextualized word embeddings approach (Devlin et al., 2019; Miaschi & Dell'Orletta, 2020; Reimers et al., 2019; Wiedemann et al., 2019). Word2Vec and GloVe are static word embeddings, meaning they do not change based on context (e.g., "table" is considered the same regardless of its context: "come to the dinner table" and "check the table of contents;" Wiedemann et al., 2019). BERT permits the embeddings to change based on context permitting greater ability to adapt to new, unseen textual topics (Devlin et al., 2019; Reimers et al., 2019). Beyond creating these contextualized word embeddings, BERT will also output its own static embeddings that have been shown to provide better predictive validities compared to Word2Vec and GloVE and Word2Vec (Fang et al., 2022; Miaschi & Dell'Orletta, 2020). Thus, BERT is one of the strongest embedding approaches for identifying trends in a corpus that is context-dependent, has a variety of topics within it, and shows strong predictive validity compared to other embedding techniques (Jia et al., 2021).

Additionally, these summary patterns found in text by NLP and word embeddings can then be compared to outside information to determine if certain features of text information are related to outside criteria. For example, showcasing the utility of BERT, Kjell et al. (2021) turned open responses to a Satisfaction with Life Scale and Harmony in Life Scale into word embeddings. The embeddings were then compared to numeric variables using a statistical model such as multiple linear regression. Using the satisfaction text-response embeddings, they predicted the rating scale scores from the Satisfaction with Life scale achieving an R^2 of .38 meaning the embeddings explained 38% of the variance within the rating scale scores. This result highlights differences in the individual responses do add to the prediction of Satisfaction with Life scale scores. Beyond predicting another variable in the same dataset, these embeddings and models can be applied to new datasets. For example, Kjell et al. (2021) used a pre-trained embedding model, based on a model by Warriner et al. (2013) that was created from almost 14,000 words that were assessed on valence, arousal and dominance, to apply valence predictions to a harmony in life scale score. BERT is superior to Word2Vec or GloVe for this purpose as it has been found to demonstrate greater content, convergent, discriminant, and predictive validity overall (Fang et al., 2022). Thus, Kjell et al. (2021) provides proof-of-concept that word embeddings can be used to predict continuous outcome variables.

The ability of NLP, and in particular word embeddings, to determine intent and sentiment has the potential to provide users additional information by further identifying patterns and trends in the text data compared with more traditional text analytics approaches. This additional information beyond text analytics assists with document classification (placing a document in the appropriate category; Pustejovsky & Stubbs, 2012). For example, NLP could assess students' written communication and sentiment to assist with categorizing students as demonstrating more growth mindset or fixed mindset. This categorization could be compared to persistence in the face of adversity (i.e., what is the student's sentiment when presented with a challenge) or high school/college retention likelihood (Hochanadel & Finamore, 2015; O'Keefe et al., 2018). The ability to create better categorizations assists the text analytics and NLP's algorithm in applying these same decision-making principles to new sources of text data.

TA, NLP, and Word Embeddings Steps and Processes

Text analytics, NLP, and word embeddings can work to identify themes in data and then be used to make an algorithm assessing differences between the themes on an outcome variable. To make a text source ready for such an algorithm, Lexaltyics describes seven computational steps, using elements of text analytics and NLP, that must occur (detailed in Table 3; Mohler, 2020). Step one involves language identification – determining in what language the document is written (e.g., English – U.S.). Selecting the appropriate language helps the algorithm appropriately break the text into categories and comprehend the sentence structure. Step two is tokenization – breaking the text into individual units of meaning one can operate on. For example, tokenization can involve dividing a sentence into single words; each word, and sometimes punctuation, would be a token (Mohler, 2020). Step three is sentence breaking - ensuring one can still tell where sentences ended within the new tokens. One can use punctuation to help determine where the sentences end, but the algorithm also needs to also interpret the words adjacent to it. For example, consider the sentence, "I went to see Dr. Adams yesterday." The algorithm must decide if the sentence ends at the period after Dr or the period after yesterday. Analyzing the tokens sequentially helps the algorithm make the correct decision. Step four is part of speech tagging – involves identifying and labeling the part of speech (e.g., noun, adjective, preposition) of each token. For example, "row" is a verb in "row your boat" and a noun in "the row of chairs." Step five is chunking – assigning part of speech tagged tokens to phrases (e.g., noun phrases, verb phrases). For example, consider the sentence, "The student was sent to the office." "The student" is the noun phrase, "was sent" is the verb phrase, and "to the office" is the prepositional phrase. Step six is syntax parsing – diagraming the sentence and determining its structure. Additionally, this is an important precursor to sentiment analysis as it can determine the sentiment of a word based on its location in the sentence diagram. Lastly, step seven is sentence chaining – connecting related sentences together based on its relative association to an overall topic; upon grouping related content together into these themes or categories, analyses can then be ran comparing them to one another leading to summative insights (Mohler, 2020).

NLP techniques, such as word embeddings, automate many of these steps with them occurring behind the scenes within the machine learning "black box" (Chen et al., 2013; Devlin et al., 2019). This section serves to help the reader understand what the predictive model is doing behind the scenes. Through this seven-step process, raw text can be grouped and categorized, helping the algorithm learn to understand named entities, track sentiment in adjective-noun combinations, and discern themes from text patterns. Additionally, this process helps eliminate irrelevant information and noise through its normalization techniques (Al-Maitah, 2018). Therefore, meaningful patterns for interpretation in qualitative data can be derived through text analytics that can aid with discerning individual and/or group differences (Liu & Huang, 2017; Nasir et al., 2020). The ability to train an algorithm to categorize content from student interactions, whether in-person or virtual, assists with scalability of SEL through being efficient schools' limited time and resources and ability to obtain results from large datasets quickly (Al-Maitah, 2018; Chen et al., 2013). Rather than having to deliver periodic self-report surveys during class, this text analytics and NLP approach can be used to continuously update its themes and classifications. Practitioners and researchers can collaborate on interpreting the significance of the themes throughout the school year for various student populations, contributing to providing flexible, low-lift approaches and solutions for SEL.

Ability to Assess Individual Differences

Text analytics can also be applied to measuring individual differences. One practical example of text analytics is a business that wants to design a new phone and understand what features customers value. The business can use text analytics on customer reviews of other phones to identify sentiment about features that were included and desires for excluded features. The frequency of certain themes in the reviews can also help the business rank-order what customers care about most. Hence, through text analytics, the computer can group together related words and identify their frequencies to deliver valuable insights to the business.

In the area of education, Liu and Huang (2017) followed a similar process to this example. Instead of reviewing customer reviews to determine phone features, they analyzed students' teaching opinion surveys to determine characteristics of students' motivation to learn. They were able to put students in groups based on the use of similar terms to describe their opinion in the survey. The groups' word choice frequency was then able to be compared to deliver insights. Ultimately, Liu and Huang (2017) were able to use text analytics to make a predictive model to group students by their motivation to learn from their previous teaching opinion survey responses. Following these examples, word embeddings can be used to assess similar terms in students' communication that correspond with their reported growth mindset, helping identify patterns and discern if any differences exist between students with different ratings of embodying growth mindset.

Another example of text analytics and NLP's application to education is how they are being increasingly incorporated in smart education (the incorporation of technology and "smart devices" into classroom instruction) due to the increased integration of technology into the classroom and the ability to collect more current feedback from students. Techniques such as sentiment analysis, opinion mining, and emotion analysis can lead to valuable insights contributing to better decision-making and an enhanced student engagement process (Mohammed et al., 2021). For example, Valcarcel et al. (2021) reviewed hundreds of thousands of reviews on RateMyTeacher.com to determine common characteristics across these descriptions of poorly rated teachers and to help drive school and district policy. Ultimately, they were able to find themes within the teacher's personality/rapport with students, duties, teaching, and assigned workload across the reviews. Specifically, they found across all reviews, 38.2% corresponded to bad teaching, 36.3% corresponded to teachers' personalities, and 11.2% corresponded to teachers as unwilling or unable to fulfill their duties. Thus, Valcarcel et al. (2021) conveyed how students' evaluative comments can generate insights that can be used to influence policy.

Text analytics has also been used to analyze the job market and advise business school curriculum. Reviewing a database of job postings, Nasir et al. (2020) used a text-analytics approach to characterize the state of the current business job market while analyzing optimal skillset themes based on different local and global job markets. These findings would aid graduate programs in re-evaluating which skills to emphasize and train across their programs. Correlating their findings of skillsets valued the most with location highlighted the supply versus demand dynamic for these skills and latent variables like compensation. Their approach also enabled them to compare how valued skillsets changed based on differences in job industry, job type, and education level. When combining all of the data together, the authors recommend professionals and business schools should focus on the following skills to convey a competitive advantage: having strong statistical foundations with a knowledge of forecasting, data mining, and visualization techniques; project and product management experience; leadership skills; initiative; and familiarity with data mining, Big Data, and machine learning toolkits (Nasir et al., 2020). Thus, their text analytics approach helped deliver practical insights for the overarching group while also delivering tailored examples for some subgroups.

As detailed earlier, there is a need to provide scalable and low-lift SEL assessment. Liu and Huang (2017) believe a data science approach is the solution. Their approach, similar to the text analytics approach previously described, ensures the quality of results can be used to solve the user's problems while providing standardization which permits faster, reliable, and manageable replication. Ultimately, incorporating text analytics can help enhance the learning experience through the insights it gains (Mohammed et al., 2021). Liu and Huang (2017) did a case study of their process assessing how university students' motivation for learning varied during a semester. Their data science approach resulted in an algorithm that enabled the university instructor to predict students' motivation to learn throughout the semester based on computer-mediated communication. Thus, as seen in these applied examples from the extant literature, text analytics results in practical predictive models enabling practitioners to make informed decisions on policy and practice (Liu & Huang, 2017; Nasir et al., 2020;

Valcarcel et al., 2021). Regarding SEL, text analytics can assist the field in predicting the intersectionality of contextual factors and their impact on exhibiting SEL competencies.

Text analytics and NLP serve as a flexible, low-lift assessment method that can analyze contextual and subgroup differences (Bailey et al., 2019; Nasir et al., 2020; Valcarcel et al., 2021). These groups can be conceptualized in numerous ways such as their classroom, which school they attend in a district, if they are in an after-school care program, and who their mentor is. Specifically, this study will discern if there are any distinctions in communication among the students grouped by their different growth mindset ratings. Furthermore, by creating a computer algorithm that can summarize data and automatically identify patterns, word embeddings permit scalability by quickly assessing and summarizing trends in student communications or open survey responses (Liu & Huang, 2017; Mohler, 2020; Valcarcel et al., 2021; Wiedemann et al., 2019). As seen in previous research, student communications can be used to assess competencies and are valuable toward influencing SEL strategies (Liu & Huang, 2017; Valcarcel et al., 2021). Thus, a word embeddings approach can help address the current limitations impacting the study of SEL and deliver valuable insights on how students exhibit growth mindset.

Current Study

This paper serves to advance the assessment of SEL by conducting an exploratory analysis of assessing a text based SEL program measurement of growth mindset. As unknown information and connectivity between various documents can be uncovered from text analytics or natural language processing, these techniques deliver predictive solutions for data analytics (Mohammed et al., 2021). This approach will also permit scalability; it can be included within e-learning platforms, passively collecting this data over manual surveys, responding to Elias's (2019) concern for the future of SEL. A conversational platform can also work to address response-sets as more rapport has been built with the individual asking questions and SEL questions can be intertwined within the conversation rather than asking the student to complete a questionnaire. Additionally, communication can occur in real time and students can actively engage with their support network. Once the SEL data is passively collected, researchers and practitioners can coordinate reviewing, discussing, and interpreting the data's meaning helping ensure practitioners are obtaining useful and actionable insights responding to some of McKown's (2019) highlighted challenges. Additionally, this paper heeds Liu and Huang's (2017) recommendation to focus on a few social-emotional skill areas before incorporating a text analytics / data science approach in large-scale education settings by focusing only on the growth mindset subdomain within the Identity SEL domain.

Previous research has suggested that text analytics and NLP, can be used to assess individual differences such as the motivation to learn and characteristics of a bad teacher (Liu & Huang, 2017; Valcarcel et al., 2021). With the increased integration of technology into the classroom and the ability to collect more current feedback from students, text analytics and NLP permits scalability and opportunities for additional evaluative techniques such as sentiment analysis, opinion mining, and emotion analysis which can lead to valuable insights contributing to better decision-making, an enhanced student engagement process, and improved policy and curriculum decisions (Mohammed et al., 2021). Since word embeddings have the potential for more integration into the classroom and current feedback from students, it can help addresses Elias's (2019) call to actions for how to feasibly scale up high-quality SEL and address the limitation of teachers/mentors not having the time/resources to appropriately rate all of their students (Ashdown & Bernard, 2011; Bailey et al., 2019).

Hence this paper presents a word embeddings evaluation of a proposed text based SEL assessment of growth mindset. This approach should address the limitations of self-report measurement and the current SEL assessment. Conducting a review of the relevant literature, to the best of our knowledge this is the first study to assess a text based SEL assessment using a word embeddings approach. Therefore, this study is explorative and will address the following research question:

> Research Question: Can analysis of text message conversations between agents and students be used to assess a student's level of growth mindset?

Methods

This project has been approved, as required, by the Institutional Review Board of Appalachian State University (February 4, 2022; IRB Reference # 110194). See Appendix A for IRB approval.

Participants

Data were collected from 815 students across the United States participating in a third-party mentoring program operated by a peer-mentoring organization that traces student academics. Full growth mindset scores were not obtained for three of the students and records for four students were duplicated, so they were dropped from supplemental analyses leaving a sample of 808 students (59.4% female, $M_{\text{age}} = 16.8$ years, $M_{\text{TimeinProgram}} = 530$ days). This mentoring program is a comprehensive digital support platform providing tutors, mentors, and advisors while emphasizing consistent, purposeful communication rather than surveys and questionnaires. Table 4 provides a summary of the student demographics.

Data was also collected from the peer-mentoring organization's agents, who digitally provide this organization's comprehensive on-demand student support services. While the specific data for the agents used in this study were not available, the peer-mentoring organization reports the average GPA of their agents is 3.63 with 36.85% of their agents currently being in college, 51.47% already having their bachelors, and 33.08% currently being in graduate school or having already received a graduate school degree. The current study included 24 agents (62.5% female, 29.2% Caucasian, M_{age} = 23.5 years, $M_{\text{TimeinProgram}}$ = 527 days) who provided growth mindset ratings on their assigned students (14-83 students per agent). The peer-mentoring organization selected all agents who have worked there for three or more years to be included. Of contacted agents, there was a 100% participation rate. Table 5 provides a summary of the agent demographics.

Text Messages

Text message conversation data was provided from students and agents involved in the program from March 18th, 2018 through February 15th, 2022. In total, 350,577 text messages were exchanged (192,660 of which were from students) during this time period among the 808 student-agent dyads ($M_{\text{TotalMessages}}$ = 433.5, $M_{\text{StudentInitatedMessages}}$ = 195.29, $M_{\text{AgentInitatedMessages}}$ = 238.21, $\mathrm{M}_{\text{NumberofTextsPerDay}}$ = 0.85). While only the message originating from the students were analyzed in the current study, Table 6 provides the descriptive statistics of the total, student-initiated, and mentor-initiated text messages. Only student-initiated texts were analyzed within the predictive models as the student's language is the primary 'source of truth' for the agents to make their growth mindset ratings. These other message characteristics are still valuable to analyze to determine if they impacted the prediction model; if significant relationships are found, future models could include texts from both students and agents.

Growth Mindset Assessment

A literature review using the RAND Education Assessment Finder (Hamilton et al., 2021) was conducted to find existing self-report Growth Mindset questions. RAND Corporation is a large non-profit research organization and reputable source. The RAND Education Assessment Finder provides information about assessments of K-12 students' interpersonal, intrapersonal, and higher-order cognitive competencies. Additionally, it permits filtering by keyword and several variables such as grade level, respondent, method of administration, administration time, item format, and fee for use (Hamilton et al., 2021).

Corresponding with this peer-mentoring organization's student population, the following filters were used: Grade Level (6-8, 9-12, postsecondary), respondent (student), method of administration (paper/pencil, digital, oral), item format (selected response, free response). This literature review resulted in 94 questions being identified from various publicly accessible assessments. Ultimately, two prompts were taken from validated, traditional self-report survey items and modified to fit into the conversational discussions that agents have with their students: "Do you think you can get smarter if you work at it?" and "Do you think it is possible to get smarter by doing challenging things?" These questions were provided to agents to bring up the growth mindset topic, if needed.

To have enough content for text analytics, it was determined that an agent's entire conversation with a student would be analyzed, and the prompts would not be required to be asked. Rather than the agents rating their students' answers to the growth mindset prompts, agents rated the extent their students display various growth mindset competencies determined from their entire communication history with the student. To facilitate agents' ratings, the rating form (see Figure 1) and training manual (see Appendix D) were created to properly train and advise agents of their responsibilities in the study. The addition of a section for agents to provide their rating rationale aids in finding trends across agents for ratings and be used to facilitate follow-up training guides for expanded SEL assessments.

As seen in Figure 1, agents were asked to rank students on four scales of growth mindset on a scale from 1 (strongly disagree) to 5 (strongly agree). These four questions align with the Harvard Taxonomy Project's self-efficacy/growth mindset subdomain (EASEL Lab, 2021a). The first prompt "This student believes that intellectual abilities and qualities can be developed and improved, rather than a skillset they are naturally stuck with" assessed the belief that one's abilities can be developed (labeled as growth mindset developed). A second evaluates if the student exhibits self-efficacy, the confidence in their ability to overcome challenges (labeled as growth mindset self-efficacy) and agents were asked to agree to the prompt "This student exhibits self-efficacy, confidence they have the ability to encounter, identify, and learn to overcome challenges they face throughout life". A third question asked agents to rate the extent "This student believes that with enough time and effort, they can overcome most challenges they come across" and gauges if a student believes with enough time and effort, they can overcome most challenges (labeled as growth mindset time and effort). The last question appraises if a student believes they have agency, the ability to make independent choices that influence their current and future life (labeled as growth mindset individual choice) and agents were asked to agree with the extent "This student believes that they have the ability to make independent choices that influence their current and future life (i.e., this student exhibits an internal locus of control rather than external)."

These four scores were then then summed to be the student's total growth mindset score (labeled growth mindset total score). The total score is comprised of these four subscales to help agents conceptually and operationally define growth-mindset. This approach also permits greater flexibility in assessing the growth mindset skills a student may exhibit in the text message conversations. Hence having four subscales and a total provides a more comprehensive assessment than just asking agents to rate a student's growth mindset based on the Dweck (2017) definition.

Collection Procedures

Student-agent conversation data was collected over text message or instant messenger via the peer-mentoring organization's digital support platform. The student-agent conversations varied in their content discussed and the amount of time between conversations. Rather than asking a pre-set list of questions in the same order, these conversations between agents and students are unstructured permitting for formation of a relationship over months of interactions and truly advising the student based on what they share with the agent. These conversations are unscripted and authentic, promoting the building of rapport and a raw exchange of information. The peer-mentoring organization has support structures and management tools in place to assess how personalized the conversation is and how relevant the resources being shared with the students are.

Agent rating data was also collected by the peer-mentoring organization. They volunteered their senior agents who have been working in the program for the past 3-4 years conveying their subject matter expert status with mentoring in the SSA program. After being selected to participate in the study, each agent was contacted individually by their supervisor to explain the purpose of the study and what was expected from the agent. Prior to filling out the spreadsheet (see Figure 1), each agent had to read through a rating form Training Manual (Appendix D provides a copy of this manual).

After receiving a custom spreadsheet from their supervisor with the names of all their students on it, each agent considered their previous conversations with the student in-relation to assessing growth mindset and its competencies. Agents were primed to consider if any previous conversation showed (an) example(s) of one of the competencies being applied and to assess if the student stopped responding when the conversation approach related to SEL (e.g., talking about a challenge and persisting through it). When making their ratings, agents were instructed to use a five-point scale where 1 corresponded to strongly disagreeing the student displays the competency and 5 corresponded to strongly agreeing the student displays the competency. For analysis, the growth mindset ratings are considered continuous; there will not be a cut-off or grouping value.

Analyses Methods

Descriptive Analysis

Before conducting any analyses using word embeddings, the distribution and intra-correlations between the growth mindset scales were examined. Next the descriptive statistics of the students, agents, and text messages were computed and examined. The correlations between the student/agent characteristics and the text messages were examined along with the relationships between the student/agent/text message characteristics and the growth mindset scale scores.

Text Data Cleaning

As with numeric data, text data also needs to be cleaned prior to analysis. Thus, before the word embedding analyses could be run, the text data needed to be refined to just student-initiated texts and consolidated. Student-initiated texts were solely used in the follow-up analyses as the content in those messages is what influences the agent's rating. The third-party partner provided data where each individual text was a separate case and a true/false column indicating if a message was student initiated. The data was subsetted to create two new variables: StudentTexts and AgentTexts. To properly get differences in embeddings by each student, all of the text messages had to be merged into one case. To accomplish this, the student text variables were grouped by a unique id number. This then resulted in a StudentTexts data set of 808 cases where each cell contained all of that student's text messages; this file was then able to be converted into embeddings.

Word Embeddings

As an exploratory analysis, the major goal of this study is to determine ‘proof of concept’ for the use to text message conversations between agents and students to assess a student’s level of growth mindset. R (R Core Team, 2021) was used to conduct the planned analyses. First, the student text messages were converted into text embeddings using the *text* package (Kjell et al., 2021) that utilizes the BERT embeddings approach (Devlin et al., 2019). BERT is a natural language framework/architecture that was pre-trained using unsupervised learning on over 2.5 billion words from the English language Wikipedia using a deep neural network. This pre-training resulted in a BERT having a “vocabulary” of over the most common 30,000 words (i.e., tokens) in the English language (McCormick & Ryan, 2019). Each token in the BERT vocabulary can be summarized by one of twenty-four layers in the neural network. Each layer is comprised of numeric values for each of BERT’s 768 dimensions which represent a variety of sematic characters about the word (Kjell et al., 2021). Interestingly, it is not yet fully understood how the various layers differ, but empirical examinations indicate that BERT’s intermediate layers (e.g., layers eight through twelve) contain sufficient linguistic information to generally yield good results (Jawahar et al., 2019; Kjell et al., 2021).

The BERT approach to converting words in the text to a summary vector has several steps. The first step is to identify the sentence beginning and ending of sentence presented in the text. BERT does this by adding special tokens (i.e., [CLS] and [SEP]) to the start and end of the respective input sentence. Next BERT tries to assign an embedding for each word in every sentence. If any word is not present in BERT’s vocabulary, BERT then uses word-piece embedding and tokenizes the word (e.g., with word "Playing" would be separated into two tokens "play" and "ing") and then assigns an embedding to the separate tokens. If a token is present in the vocabulary, then BERT tries to break the word into the largest possible token contained in the vocabulary, and as a last resort will decompose the word into individual characters. It is because of this automated feature that BERT can always represent a word as, at the very least, the collection of its individual characters (McCormick & Ryan, 2019).

After the embeddings (i.e., numeric vector of meaning) are assigned to every token (i.e., term in BERT's pre-determined vocabulary) in a corpus (i.e., a structured collection of texts), the embeddings of tokens are combined or aggregated into one word embedding to sentences or paragraphs. This can be achieved by taking the mean, minimum or maximum value of each dimension (i.e., meaning of the vector – gender-ness, tense, sentiment) of the embeddings. The default method in the *text* package is to use the mean of the vectors to create one embedding for each case in a corpus. The *text* package also allows you to return the results of multiple layers. In the current study, layers 11 and 12 were concatenated, which resulted in each student's set of text being represented by a vector of 1536 dimensions.

The *textEmbed* function resulted in all categorical/text variables in the relevant dataframe being converted into contextualized embeddings, saved in the output as the same name. Word embeddings create an ordered vector, a list of values, that depicts the meaning of a word (e.g., .1849, .0692, .0427, .00992). Naturally, this current output is meaningless to the coder; however, it makes sense to the machine learning the algorithm. Kjell et al. (2021) summarizes, "the numbers may be seen as coordinates in a geometric space that comprises several hundred dimensions (a high dimensional space). The closer positioned two words are in this space (i.e., the more similar their vector embeddings are), the more similar the words are in meaning" (p. 7).

The dimension output does not provide context as to what that specific dimension represents, but these elements can represent topics such as gender-ness of a word, a created theme (e.g., growth mindset), or tense. All of these dimensions are used to help create and refine the predictive model. Hence, the *text* package permits access to high-quality pre-trained embeddings that have been created from numerous previous analyses to enhance the quality of embeddings for this novel dataset while also permitting access to contextualized embedding options.

Predictive Modeling

Once embeddings were created, relationships were then examined between the word embeddings and growth mindset scale scores using the *textTrainLists* function. Within this function, principal component analysis (PCA) and ridge regression with cross-validation were used to create a predictive model for each of the growth mindset scores using the resulting 1536 dimensions from the word embeddings (Kjell et al., 2021). PCA was a pre-processing step in the *textTrainLists* function that served as a means to accurately report and evaluate a large number of variables using fewer components, while still preserving the dimensions of the data. It is widely described as a data reduction method used to summarize a large set of variables. The default in *textTrainLists* is the "min_halving", which is a function that selects the number of PCA components based on number of participants and feature (word embedding dimensions) in the data. In this case, the PCA retained 768 components for the ridge regression analysis[1].

These components were then entered into a ridge regression (Hoerl & Kennard, 1970) which follows a similar process as multiple linear regression, but it specializes on data that suffers from multicollinearity, the extent to which independent variables are correlated and have a joint impact. Additionally, it is used when there are a large number of predictors leading to concerns of the data in the current situation to remain applicable to other situations. This regression uses beta weights to regularize the coefficients by penalizing larger coefficients helping shrink them closer to the “true” population parameters. This regularization and shrinkage help tune the model which reduces its bias.

Cross validation is used to tune the ridge regression and prediction model. Cross validation helps choose the statistical model with minimal extraneous error included while assessing its predictive ability and generalizability (Kjell et al., 2021). Specifically, a K-fold approach was used to randomly divide the data set into a training set and a testing set. Within k-fold the “k” refers to the number of groups created; the default was 10. For each group, a prediction model was estimated using the training set and then applied to the testing set, creating predicted values. These predicted values were then compared to the observed values and residuals could be calculated. The average of all the folds serves as the performance metric for the model; thus, this approach helps reduce bias (Kjell et al., 2021).

[1] The min_halving formula was that the number of components to be retained would be the max of two possible sets of numbers: 1) the smaller of either (number_features/2) or number_participants/2); and 2) the smaller of either 50 or the number of model features.

Ultimately, *textTrainLists* created a variety of useful outputs. First, it listed all predicted values for each case. It then reported the finalized prediction model for each dependent variable. Metrics for the model were also provided such as a Pearson's product-moment correlation. This is the correlation between the predicted value from the model and the actual observed value and is equivalent to an R value from a regression model. Squaring this value will provide R^2 which provides the amount of variance explained by the model detailing how well it fits the observed data.

To further understand the accuracy of these prediction models and the R values, root mean square error (RMSE) values were calculated. This value shows the standard deviation of the residuals for each prediction model. The lower the RMSE value, the better a given model can "fit" a dataset, having the predicted values be as close to the observed values. However, the context of the data one is working with determines what truly is a "low" value. Therefore, one can normalize their RMSE by taking it and dividing it by the difference between its maximum and minimum value (RMSE / [max -min]); the closer the normalized RMSE is to zero, the better a given model is able to "fit" a dataset.

Additionally, the distribution of residuals and homoscedasticity, the assumption of similar variance across all groups being compared indicating equivalent prediction accuracy across the entire data, was assessed for each prediction model. Histograms of the unstandardized residuals and scatterplots with the residual value on the y-axis and fitted value on the x-axis were used for these evaluations.

Bias Check

The prediction model outcomes only revealed how well the model overall predicted growth mindset score ratings. It alone informed if it predicted better for one subgroup over another. Therefore, the residuals, the difference between the actual and predicted value from the embeddings, were used to assess if the prediction model was more or less accurate based on student, agent, or text message characteristics. To assess for differences by student gender, agent gender, and agent ethnicity, one-way ANOVAs were conducted due to having a categorical predictor and continuous outcome. If the ANOVA indicated a significant difference by having a significant p-value and meaningful eta-squared, post-hoc analyses were run to determine where the significant difference was amongst the categorical groups. Specifically, a Bonferroni post-hoc test was run due to the large sample size, low variability, and unknown expected effect size. It is not as restrictive as other post-hoc tests and does not overly increase or reduce the likelihood of type I or type II errors as other post-hoc test comparisons. Additionally, the Cohen's d effect size was also provided to determine the size of the difference between groups. Correlations were run to assess the relationship between each model's residuals and student/agent age, student/agent length of time in the program, total number of messages sent, number of student/agent-initiated messages, average number of texts per day, and percent of student/agents messages sent.

Ultimately, these results can help determine differences in accuracy by analyzing the post-hoc comparisons to determine where differences occur. This output informs if the growth mindset scores were being skewed and can advise follow-up training if it is a mentor rating bias or advise future models to control more for certain demographic variables. Additionally, these results help answer the research question by providing caveats to the model's overall predication accuracy.

Results

Growth Mindset Scale Descriptives

The descriptive statistics for the student/agent/text message characteristics are described within the method section and detailed in Tables 4-6. The 24 agents provided ratings for 808 students on five growth mindset scales. Figures 2 through 6 show the histograms for each scale. The higher end of the spectrum (4-5 or 15-20 for the total score scale) had the greatest frequency of scores, and there were few scores on the lower end of the spectrum (1-2 or 1-10 for the total score scale). Examining these histograms reveal all growth mindset scales were negatively skewed. Furthermore, the descriptive statistics table (see Table 7) further highlights this skew by showing the median score on the growth mindset total score scale was a 16, indicating at least 440 students scored a 16 or higher on the 20-point scale. Thus, having so few cases at the lower end of the spectrum may reduce the models' accuracy for predicting the growth mindset score due to range restriction, highlighting the need to assess variables that may further bias the predictive validities.

Correlations examining the relationships between the four growth mindset scores were conducted (see Table 8). Predictably, all scales were strongly correlated with one another (r = .81 - .89). All scales are spuriously inflated with the total growth mindset scores as they are part-whole correlations. These strong intercorrelations indicated the agents appeared to rate the each of the growth mindset scales in a similar way for a student.

Student/Agent/Text Characteristics and Growth Mindset Scale Correlations

Correlations examined the relationship between student/mentor characteristics with text message characteristics (see Table 9) and with growth mindset scores (see Table 10). The relationship between student/mentor characteristics with text message characteristics indicated that student age had a positive relationship with student time in the program (r = .25), total messages (r = .12), number of student-initiated messages (r = .09), and number of mentor-initiated messages (r = .15). These results indicate the older a student is, the longer on-average they have been in the program, and the greater number of total, student-initiated, and mentor-initiated messages were exchanged. Furthermore, student age had a significant positive correlation with all growth mindset scales: developed (r = .15), self-efficacy (r = .18), time and effort (r = .18), independent choice (r = .17), and total score (r = .18). This result indicates the older a student is, the higher their growth mindset score was on each scale. Thus, these correlations highlight student age should be assessed for biases.

Student time in the program had a significant positive correlation with agent time in the program (r = .27), total messages (r = .40), number of student-initiated messages (r = .30), and number of mentor-initiated messages (r = .49). These results indicate the longer a student has been in the program, the longer on-average the agent had also been in the program, and the greater number of total, student-initiated, and mentor-initiated messages were exchanged. Furthermore, student time in the program had a significant positive correlation with all growth mindset scales except for developed: self-efficacy (r = .08), time and effort (r = .08), independent choice (r = .08), and total score (r = .07). This result indicates the longer a student has been in the program, the higher their growth mindset score was on each scale besides growth mindset developed. Thus, these correlations highlight student time in the program should be assessed for biases.

Agent age had a negative correlation with agent time in the program (r = -.10), total messages (r = -.07), and number of mentor-initiated messages (r = -.09). These results indicate the older an agent is, the shorter on-average they have been in the program, and the lower number of total and mentor-initiated messages were exchanged, with fewer texts exchanged per day. Furthermore, there were no significant correlations with the growth mindset scales. Thus, these correlations highlight agent age should be assessed for biases.

Agent time in the program had a significant positive correlation with total messages (r = .09) and number of mentor-initiated messages (r = .14). These results indicate the longer an agent has been in the program, the higher number of total and mentor-initiated messages were exchanged. Furthermore, agent time in the program had a significant positive correlation with all growth mindset scales except for developed: self-efficacy (r = .14), time and effort (r = .09), independent choice (r = .09), and total score (r = .10). This result indicates the longer an agent has been in the program, the higher their growth mindset rating was on each scale besides growth mindset developed. Thus, these correlations highlight agent length of time in the program should be assessed for biases.

Prediction Models

The student text messages were run through the BERT embeddings model resulting in each student having 1536 feature vectors representing the semantic meaning of their text messages. These dimension vectors were then submitted into the word embeddings prediction model to determine the extent to which these numeric representations of student texts predicted each of the study's five agent-provided growth mindset ratings. The model fit statistics from the prediction models are presented in Table 11.

Growth Mindset Developed

The prediction model using the BERT text embeddings to predict growth mindset developed was statistically significant, R = .40, $p < .001$ (see Table 11), indicating patterns that were detected within the student text messages were able to predict the growth mindset developed rating score. Comparing the predictive model's error fit indices (RMSE = 0.96, normalized RMSE = .24) to the naïve model's error fit indices ($RMSE_{Null} = 1.05$, normalized $RMSE_{Null} = .26$) revealed that the prediction model fit the data better than a null prediction model (i.e., the average of the growth mindset developed score).

The prediction model diagnostics indicated that the model fit the data reasonably well. Examining a histogram of the prediction model's residuals revealed that residuals were slightly negatively skewed (see Figure 7) indicating that the model might be more prone to overpredict growth mindset developed scores (i.e., predicted values were larger than actual scores). Looking at the residuals by fitted values plot (see Figure 8) revealed that the residuals mostly met the assumption of homogeneity. The model was slightly more likely to underpredict (i.e., positive residuals) when predicting lower values and more likely to overpredict (i.e., negative residuals) when predicting higher values.

Follow-up analyses on prediction model residuals were conducted to determine if the model's predictive accuracy differed by reported student, mentor, and text characteristics. One-way ANOVAs were conducted to assess student gender, agent gender, and agent ethnicity (see Table 12). There was no difference in model accuracy between male and female students, $F(1,806)= 1.44$, $p = .23$, $\eta^2 = .002$. There was also no difference in model accuracy between male and female agents, $F(1,806)= 1.12$, $p = .29$, $\eta^2 = .001$.

A significant difference in model accuracy was observed for agent ethnicity, $F(7,800)= 4.77$, $p < .001$, $\eta^2 = .04$. The Bonferroni post-hoc analysis indicated three significant group differences: between Caucasian and African American agents, $t(800) = 4.83$, $p < .001$, $d = -.59$; Caucasian and Eastern European agents, $t(800) = -4.02$, $p < .01$, $d = -.53$; and Caucasian and Indian agents, $t(800) = -3.41$, $p < .05$, $d = -.44$. Post hoc estimated marginal means can be found in Table 13.

A series of correlations were also conducted to assess student and agent age, time in the program, and amount of student and mentor messages exchanged (see Table 14). The results revealed student age significantly correlated with the growth mindset developed score's residual, $r = .08$, $p < .05$. As student age increased, the actual value was larger than the predicted value. Thus, the significant, positive correlation indicates the prediction model tended to under predict the growth mindset developed score for older students. There was no difference in model residuals by agent age ($p = .579$), student time in the program ($p = .276$), or agent time in the program ($p = .150$). There were also no significant relationships between quantity of messages exchanged, whether total ($p = .226$), student-initiated ($p = .485$), mentor-initiated ($p = .073$), or average number of messages per day ($p = .923$) and the model residuals. However, the relative proportion of who initiated sending messages did result in a significant difference in model residuals; specifically, student-initiated percent, $r = .18$, $p < .001$, and mentor-initiated percent $r = -.18$, $p < .001$, had statistically significant relationships with the model residuals. These relationships indicated that as students sent a larger proportion of the total text messages, the residual tended to be more positive indicating the model tended to underpredict their growth mindset development ratings.

Growth Mindset Self-Efficacy

The prediction model using the BERT text embeddings to predict growth mindset developed was statistically significant, R = .37, $p < .001$ (see Table 11), indicating patterns that were detected within the student text messages were able to predict the growth mindset self-efficacy rating score. Comparing the predictive model's error fit indices (RMSE = 1.01, normalized RMSE = .25) to the naïve model's error fit indices ($RMSE_{Null} = 1.09$, normalized $RMSE_{Null} = .27$) revealed that the prediction model fit the data better than a null prediction model (i.e., the average of the growth mindset self-efficacy score).

The prediction model diagnostics indicated that the model fit the data reasonably well. Examining a histogram of the prediction model's residuals revealed that residuals were slightly negatively skewed (see Figure 9) indicating that the model might be more prone to overpredict growth mindset self-efficacy scores. Looking at the residuals by fitted values plot (see Figure 10) revealed that the residuals mostly met the assumption of homogeneity. The model was slightly more likely to underpredict (i.e., positive residuals) when predicting lower values and more likely to overpredict (i.e., negative residuals) when predicting higher values.

Follow-up analyses on prediction model residuals were conducted to determine if the model's predictive accuracy differed by reported student, mentor, and text characteristics. One-way ANOVAs were conducted to assess student gender, agent gender, and agent ethnicity (see Table 15). There was no difference in model accuracy between male and female students, $F(1,806)= .53$, $p = .47$, $\eta^2 = .001$. There was also no difference in model accuracy between male and female agents, $F(1,806)= .495$, $p = .48$, $\eta^2 = .001$.

A significant difference in model accuracy was observed for agent ethnicity, $F(7,800)= 3.49$, $p < .001$, $\eta^2 = .03$. The Bonferroni post-hoc analysis indicated one significant group difference between Indian and Western European agents, $t(800) = 3.16$, $p < .05$, $d = 0.60$. Post hoc estimated marginal means can be found in Table 16.

A series of correlations were also conducted to assess student and agent age, time in the program, and amount of student and mentor messages exchanged (see Table 17). The results revealed student age significantly correlated with the growth mindset self-efficacy score's residual, $r = .12$, $p < .01$. As student age increased, the actual value was larger than the predicted value. Thus, the significant, positive correlation indicates, the prediction model tended to under predict the growth mindset self-efficacy score for older students.

Additionally, results revealed student time in the program, $r = .09$, $p < .01$, and agent time in the program, $r = .13$, $p < .001$, significantly correlated with the growth mindset self-efficacy score's residual. As a student or agent spent more time in the program, the actual value was larger than the predicted value. Thus, the significant, positive correlation indicates as the students and agents spent more time in the program, the prediction model tended to under predict the growth mindset self-efficacy score.

Furthermore, results revealed quantity of agent-initiated messages, $r = .08$, $p < .05$, significantly correlated with the growth mindset self-efficacy score's residual. As an agent sent more messages, the actual value was larger than the predicted value. Thus, the significant, positive correlation indicates as agents sent more messages, the prediction model tended to under predict the growth mindset self-efficacy score.

There was no difference in model residuals by agent age ($p = .190$), total messages exchanged ($p = .169$), student-initiated messages ($p = .470$), or average number of student-sent messages per day ($p = .923$) and the model residuals. However, the relative proportion of who initiated sending messages did result in a significant difference in model residuals; specifically, student-initiated percent, $r = .18$, $p < .001$, and mentor-initiated percent, $r = -.18$, $p < .001$, had statistically significant relationships with the model residuals. These relationships indicated that as students sent a larger proportion of the total text messages, the residual tended to be more positive indicating the model tended to underpredict their growth mindset self-efficacy ratings.

Growth Mindset Time and Effort

The prediction model using the BERT text embeddings to predict growth mindset time and effort was statistically significant, R = .41, $p < .001$ (see Table 11), indicating patterns that were detected within the student text messages were able to predict the growth mindset time and effort rating score. Comparing the predictive model's error fit indices (RMSE = .96, normalized RMSE = .24) to the naïve model's error fit indices ($RMSE_{Null} = 1.05$, normalized $RMSE_{Null} = .29$) revealed that the prediction model fit the data better than a null prediction model (i.e., the average of the growth mindset time and effort score).

The prediction model diagnostics indicated that the model fit the data reasonably well. Examining a histogram of the prediction model's residuals revealed that residuals were slightly negatively skewed (see Figure 11) indicating that the model might be more prone to overpredict growth mindset time and effort scores. Looking at the residuals by fitted values plot (see Figure 12) revealed that the residuals mostly met the assumption of homogeneity. The model was slightly more likely to underpredict (i.e., positive residuals) when predicting lower values and more likely to overpredict (i.e., negative residuals) when predicting higher values.

Follow-up analyses on prediction model residuals were conducted to determine if the model's predictive accuracy differed by reported student, mentor, and text characteristics. One-way ANOVAs were conducted to assess student gender, agent gender, and agent ethnicity (see Table 18). There was no difference in model accuracy between male and female students, $F(1,806)= .003$, $p = .96$, $\eta^2 < .001$. There was also no difference in model accuracy between male and female agents, $F(1,806)= .003$, $p = .95$, $\eta^2 < .001$.

A significant difference in model accuracy was observed for agent ethnicity, $F(7,800) = 6.67$, $p < .001$, $\eta^2 = .06$. The Bonferroni post-hoc analysis indicated six significant group differences: between African American and Asian agents, $t(800) = 4.23$, $p < .001$, $d = .5$; African American and Caucasian agents, $t(800) = 5.42$, $p < .001$, $d = .66$; African American and Western European agents, $t(800) = 3.31$, $p < .05$, $d = .61$; Caucasian and Eastern European agents, $t(800) = -3.22$, $p < .05$, $d = -.43$; Caucasian and

Hispanic agents, $t(800) = -3.18$, $p < .05$, $d = -.51$; and Caucasian and Indian agents, $t(800) = -2.96$, $p < .01$, $d = -.54$. Post hoc estimated marginal means can be found in Table 19.

A series of correlations were also conducted to assess student and agent age, time in the program, and amount of student and mentor messages exchanged (see Table 20). The results revealed student age significantly correlated with the growth mindset time and effort score's residual, $r = .18$, $p < .01$. As student age increased, the actual value was larger than the predicted value. Thus, the significant, positive correlation indicates the prediction model tended to under predict the growth mindset time and effort score for older students.

Additionally, results revealed student time in the program, $r = .08$, $p < .05$, and agent time in the program, $r = .09$, $p < .05$, significantly correlated with the growth mindset time and effort score's residual. As a student or agent spent more time in the program, the actual value was larger than the predicted value. Thus, the significant, positive correlation indicates as the students and agents spent more time in the program, the prediction model tended to under predict the growth mindset time and effort score. There was no difference in model residuals by agent age ($p = .380$), total messages exchanged ($p = .725$), student-initiated messages ($p = .877$), mentor-initiated messages ($p = .239$), or average number of student-sent messages per day ($p = .656$) and the model residuals. However, the relative proportion of who initiated sending messages did result in a significant difference in model residuals; specifically, student-initiated percent, $r = .16$, p < .001, and mentor-initiated percent, $r = -.16$, p < .001, had statistically significant relationships with the model residuals. These relationships indicated that as students sent a larger proportion of the total text messages, the residual tended to be more positive indicating the model tended to underpredict their growth mindset time and effort ratings.

Growth Mindset Independent Choice

The prediction model using the BERT text embeddings to predict growth mindset independent choice was statistically significant, R = .40, $p < .001$ (see Table 11), indicating patterns that were detected within the student text messages were able to predict the growth mindset independent choice rating score. Comparing the predictive model's error fit indices (RMSE = 1.00, normalized RMSE = .25) to the naïve model's error fit indices ($RMSE_{Null} = 1.09$, normalized $RMSE_{Null} = .27$) revealed that the prediction model fit the data better than a null prediction model (i.e., the average of the growth mindset independent choice score).

The prediction model diagnostics indicated that the model fit the data reasonably well. Examining a histogram of the prediction model's residuals revealed that residuals were slightly negatively skewed (see Figure 13) indicating that the model might be more prone to overpredict growth mindset independent choice scores. Looking at the residuals by fitted values plot (see Figure 14) revealed that the residuals mostly met the assumption of homogeneity. The model was slightly more likely to underpredict (i.e., positive residuals) when predicting lower values and more likely to overpredict (i.e., negative residuals) when predicting higher values.

Follow-up analyses on prediction model residuals were conducted to determine if the model's predictive accuracy differed by reported student, mentor, and text characteristics. One-way ANOVAs were conducted to assess student gender, agent gender, and agent ethnicity (see Table 21). There was no difference in model accuracy between male and female students, $F(1,806) = .003$, $p = .95$, $\eta^2 < .001$. There was also no difference in model accuracy between male and female agents, $F(1,806) = .01$, $p = .92$, $\eta^2 < .001$.

A significant difference in model accuracy was observed for agent ethnicity, $F(7,800) = 5.67$, $p < .001$, $\eta^2 = .05$. The Bonferroni post-hoc analysis indicated three significant group differences: between Caucasian and African American agents, $t(800) = 4.23$, $p < .001$, $d = -.55$; Caucasian and Eastern European agents, $t(800) = -3.47$, $p < .05$, $d = -.46$; and Caucasian and Indian agents, $t(800) = -4.12$, $p < .01$, $d = -.54$. Post hoc estimated marginal means can be found in Table 22.

A series of correlations were also conducted to assess student and agent age, time in the program, and amount of student and mentor messages exchanged (see Table 23). The results revealed student age significantly correlated with the growth mindset independent choice score's residual, $r = .10$, $p < .01$. As student age increased, the actual value was larger than the predicted value. Thus, the significant, positive correlation indicates the prediction model tended to under predict the growth mindset independent choice score for older students.

Additionally, results revealed student time in the program, $r = .07$, $p < .05$, and agent time in the program, $r = .07$, $p < .05$, significantly correlated with the growth mindset independent choice score's residual. As a student or agent spent more time in the program, the actual value was larger than the predicted value. Thus, the significant, positive correlation indicates as the students and agents spent more time in the program, the prediction model tended to under predict the growth mindset independent choice score.

Furthermore, results revealed quantity of agent-initiated messages, $r = .09$, $p < .01$, significantly correlated with the growth mindset independent choice score's residual. As an agent sent more messages, the actual value was larger than the predicted value. Thus, the significant, positive correlation indicates as agents sent more messages, the prediction model tended to under predict the growth mindset independent choice score.

There was no difference in model residuals by agent age ($p = .270$), total messages exchanged ($p = .053$), student-initiated messages ($p = .195$), or average number of student-sent messages per day ($p = .903$) and the model residuals. However, the relative proportion of who initiated sending messages did result in a significant difference in model residuals; specifically, student-initiated percent, $r = .17$, $p < .001$, and mentor-initiated percent, $r = -.17$, $p < .001$, had statistically significant relationships with the model residuals. These relationships indicated that as students sent a larger proportion of the total text messages, the residual tended to be more positive indicating the model tended to underpredict their growth mindset independent choice ratings.

Growth Mindset Total

The prediction model using the BERT text embeddings to predict growth mindset total was statistically significant, R = .43, $p < .001$ (see Table 11), indicating patterns that were detected within the student text messages were able to predict the growth mindset total rating score. Comparing the predictive model's error fit indices (RMSE = 3.60, normalized RMSE = .23) to the naïve model's error fit indices ($RMSE_{Null} = 3.99$, normalized $RMSE_{Null} = .25$) revealed that the prediction model fit the data better than a null prediction model (i.e., the average of the growth mindset total score).

The prediction model diagnostics indicated that the model fit the data reasonably well. Examining a histogram of the prediction model's residuals revealed that residuals were slightly negatively skewed (see Figure 15) indicating that the model might be more prone to overpredict growth mindset total scores. Looking at the residuals by fitted values plot (see Figure 16) revealed that the residuals mostly met the assumption of homogeneity. The model was slightly more likely to underpredict (i.e., positive residuals) when predicting lower values and more likely to overpredict (i.e., negative residuals) when predicting higher values.

Follow-up analyses on prediction model residuals were conducted to determine if the model's predictive accuracy differed by reported student, mentor, and text characteristics. One-way ANOVAs were conducted to assess student gender, agent gender, and agent ethnicity (see Table 24). There was no difference in model accuracy between male and female students, $F(1,806) = 1.24$, $p = .27$, $\eta^2 = .002$. There was also no difference in model accuracy between male and female agents, $F(1,806) = .26$, $p = .611$, $\eta^2 < .001$.

A significant difference in model accuracy was observed for agent ethnicity, $F(7,800) = 5.94$, $p < .001$, $\eta^2 = .05$. The Bonferroni post-hoc analysis indicated five significant group differences: between African American and Asian agents, $t(800) = 3.32$, $p < .05$, $d = .39$; African American and Caucasian agents, $t(800) = 5.00$, $p < .001$, $d = .61$; African American and Western Europe agents, $t(800) = 3.23$, $p < .05$, $d = .60$; Caucasian and Eastern European agents, $t(800) = -3.54$, $p < .05$, $d = -.47$; and Caucasian and Indian agents, $t(800) = -4.02$, $p < .01$, $d = -.52$. Post hoc estimated marginal means can be found in Table 25.

A series of correlations were also conducted to assess student and agent age, time in the program, and amount of student and mentor messages exchanged (see Table 26). The results revealed student age significantly correlated with the growth mindset total score's residual, $r = .09$, $p < .05$. As student age increased, the actual value was larger than the predicted value. Thus, the significant, positive correlation indicates the prediction model tended to under predict the growth mindset total score for older students.

Additionally, results revealed student time in the program, $r = .08$, $p < .05$, and agent time in the program, $r = .09$, $p < .05$, significantly correlated with the growth mindset total score's residual. As a student or agent spent more time in the program, the actual value was larger than the predicted value. Thus, the significant, positive correlation indicates as the students and agents spent more time in the program, the prediction model tended to under predict the growth mindset total score.

Furthermore, results revealed quantity of agent-initiated messages, $r = .08$, $p < .05$, significantly correlated with the growth mindset total score's residual. As an agent sent more messages, the actual value was larger than the predicted value. Thus, the significant, positive correlation indicates as agents sent more messages, the prediction model tended to under predict the growth mindset total score.

There was no difference in model residuals by agent age ($p = .259$), total messages exchanged ($p = .169$), student-initiated messages ($p = .470$), or average number of student-sent messages per day ($p = .770$) and the model residuals. However, the relative proportion of who initiated sending messages did result in a significant difference in model residuals; specifically, student-initiated percent, $r = .18$, $p < .001$, and mentor-initiated percent, $r = -.18$, $p < .001$, had statistically significant relationships with the model residuals. These relationships indicated that as students sent a larger proportion of the total text messages, the residual tended to be more positive indicating the model tended to underpredict their growth mindset total ratings.

Discussion

Heeding the call for innovation in assessment of SEL (Ashdown & Bernard, 2011; Bailey et al., 2019; Jones & Bouffard, 2012; Liu & Huang, 2017), this study explored the feasibility of using word embeddings of student-agent text conversations to assess a student's growth mindset. Agents provided ratings of their students' growth mindset on four scales (developed, independent choice, self-efficacy, and time and effort) that were summed to be the student's total growth mindset score. The full text conversation history between students and agents were collected and turned into word embeddings utilizing the *text* package (Kjell et al., 2021) and BERT (Devlin et al., 2019). The *text* package then created predictive models of the word embeddings predicting the continuous growth mindset scores. Follow-up analyses on student, agent, and text characteristics were then conducted to assess if they biased the prediction accuracy of the model.

The results indicated all prediction models were significant (predictive validities between .37 and .43), supporting the research question that analysis of text message conversations between agents and students can be used to assess a student's level of growth mindset. The growth mindset total score had the greatest prediction validity. These models can explain between 14% and 19% of the variance in the scale score. Thus, these results indicate text analytics and natural language processing can be used to assess growth mindset and perhaps be expanded to other SEL assessment.

With continued training of the models on more tailored, pre-trained word embeddings, the model's predictability is expected to increase (Alsentzer et al., 2019; Wiedemann et al., 2019). Further, this research supports the scalability of this approach – with more data and training across time, the predictive model is expected to improve over time and be able to appropriately predict based on its previous trained embeddings rather than requiring a respondent to report growth mindset scores. Thus, with enough iterations the trained embeddings can serve as a comparable criterion rather than requiring respondent-reported or self-reported scores. This ability to use its previous embeddings will assist instructors in not sacrificing valuable instructional time for self-report assessments (Ashdown & Bernard, 2011; Bailey et al., 2019; Jones & Bouffard, 2012). Additionally, this iterative and ongoing process can assist with a flexible intervention approach that also acknowledges environmental and contextual effects on SEL competencies, hence continued research on factors that promoted the model to overpredict or underpredict student SEL levels.

The RMSE values were adequate for all models but indicated it would be valuable to conduct follow-up analyses to determine if the prediction models were biased by other variables. Follow-up analyses indicated prediction model residuals varied by student/agent age, student/agent time in the program, agent ethnicity, number of agent-initiated messages, and relative proportion of who initiated sending messages. Student age impacted all five models while student time in the program impacted all models beyond growth mindset developed. Consistently, there was a significant positive relationship between student age and the model's score residuals indicating under prediction for older students. This was a surprising finding given the literature describing growth mindset typically increases monotonically throughout elementary and middle school before levelling off in high school (West et al., 2020). With the mean age of the sample being 16.8, an age commonly associated with 10th or 11th grade, it was expected that growth mindset would level off. This may speak to the model believing growth mindset would increase over time and not have a curvilinear, leveling-off pattern. Additionally, there was a significant positive correlation between student time in the program and the models' score residuals indicating the model tended to underpredict for students who had been in the program longer. This finding may be explained by the significant correlation with student age (r = .25) or perhaps an attrition bias of students dropping out of the program or reducing their participation with their agent/mentor.

Each model found at least one significant difference between its residuals and agent ethnicity. A common difference across models was between Caucasian and African American agents which indicated Caucasian agents tended to be less lenient with their ratings hence the model tended to over predict their ratings (a negative mean value) while African American agents tended to be more lenient with their ratings hence the model tended to under predict their rating (a positive mean value). These significant differences could possibly be explained by the mentors having different levels of growth mindset themselves; however, previous research indicates race/ethnicity differences in growth mindset narrow by 12th grade with all races/ethnicities seeing an increase in growth mindset after (West et al., 2020). It should also be noted the low sample size; the largest ethnicity, Caucasian, had a sample size of 7, so ethnicity differences in this study may not be generalizable.

The relative proportion of who initiated sending messages resulted in a significant difference in each model's residuals; when students sent a larger proportion of the total text messages, the model tended to underpredict their growth mindset ratings. This was a surprising finding given the students were more actively involved in the conversation; however, this may be explained by the quality of text communication. The agent and student may have talked a lot – about topics not related to growth mindset, or about topics that provided the agent with more examples of lower growth mindset categorizations.

Ultimately, text analytics and natural language process have been shown to feasibly assess students' growth mindset. If a factor is found to overpredict or underpredict, a group of researcher and practitioners can convene to determine what changes need to be made. This approach can aid high-impact SEL assessment; given text analytics and NLP is an iterative process, it provides the opportunity for researchers and practitioners to review, discuss, and interpret the prediction models' meaning to decide what to do regarding curriculum and/or targeted interventions (McKown, 2019; Mohammed et al., 2021). This study could be enhanced by also assessing the SEL (in this study's case, growth mindset specifically) of the individuals (e.g., teacher/mentor) providing the SEL ranking. Teachers' comfort with SEL influences classroom instruction and ultimately academic achievement and student engagement outcomes (Yang et al., 2018). Additionally, Collie et al. (2012) assessed how teachers' perceptions of SEL and school climate influenced their sense of stress, teaching efficacy, and job satisfaction. A teacher's comfort in implementing SEL moderately correlated (r = .34) with teaching efficacy, weakly correlated with stress related to student behavior (r = -.21), and weakly correlated with job satisfaction (r = .18). Furthermore, a teacher's commitment to improving their SEL skills weakly correlated with stress related to student behavior (r = .23), weakly correlated with workload stress (r = .24), and weakly correlated with job satisfaction (r = .11). Therefore, future studies could also assess the instructor's SEL in addition to the students as they are a crucial factor in a supportive environment for SEL.

Additionally, there are numerous existing text analytics techniques; different techniques may be better for this type of analysis depending upon the applied study situation and the resources available to the researcher (Al-Maitah, 2018; Liu & Huang, 2017; Nasir et al., 2020; Valcarcel et al., 2021). Thus, the same data can be used in different packages and modeling tools to see if there are any differences in results based upon methodology. Research as a whole would benefit from more replication (plus extension) studies.

One such tool future studies could use is Prodigy (2021), an annotation tool linked to Python that assists with conducting error analysis and training and evaluating data for machine learning modules. Specifically, its named entity recognition (tagging names, concepts, or key phrases that do not overlap), span categorization (extracting longer phrases and nested expressions that may overlap), and text classification (automatically applying labels and groupings to the text) features will be utilized. Its user interface (UI) can automatically pre-tokenize the data, knowing where words begin and end, letting analysis begin at the part-of-speech-tagging stage of preparing raw data for text analytics and NLP. After enough data has been labeled, Prodigy permits to train the tool to automatically start labeling other data following similar logic (Prodigy, 2021).

Another approach could be utilizing decision trees or random forests to predict groupings of students. Those types of models have been shown to (generally) provide better prediction of outcomes than regression (Fang et al., 2022). Thus, if the growth mindset scale scores can be categorized into ‘low’ and ‘high’ categorizations, these models can assist in finding trends in language between the two groups perhaps improving overall model predictive validities. However, the present study had negative skew in all of the growth mindset scale scores with few scores at the lower end of the spectrum. This skew inhibited decision trees and random forests from being run due to very few “low” growth mindset scores reducing the ability to compare a “low” and “high” categorization.

Ultimately, this skew contributed to range restriction which may be explained by two factors in this study. First, this study population is comprised of students who opted-in to this third-party service to be assigned an agent/mentor; there may be a difference in growth mindset scores between the students that do and do not opt-in to this program. Second, this skew may indicate agents were too lenient with their ratings. Attempts were made to inform agents that both high and low scores were beneficial; hence a quote from the training manual (see Appendix D), “a high or low score is not necessarily better. Accurately rating each student can help align follow-up interventions and support structures to benefit the student’s long-term development.” Perhaps this manual can be further refined to make this more salient as well as emphasizing these results will not be used in relation to an agent’s performance evaluation or job compensation.

Contributions and Implications

This study offers several contributions regarding research in the areas of SEL and text analytics. While there is a growing focus on SEL by practitioners and researchers as well as continued focus on text analytics processes, this is the first study to attempt a word embedding assessment approach for growth-mindset. The created models indicates that word embeddings of student-agent text conversations can be used to assess a student's growth-mindset, achieving predictive validities between .37 and .43 and thus extending the research by reporting another assessment approach beyond self-report surveys. However, further research is needed to further refine these predictive models and continue to assess this approach's feasibility.

Another contribution of this research is the practical implications it will have for SEL practitioners. Given the continued integration of smart technology into the classrooms (Mohammed et al., 2021), SEL assessment can expand beyond self-report surveys to include analysis of trends in students' writing and participation in discussion boards. This can help get a more well-rounded understanding of a student's growth mindset. Additionally, by providing more opportunities and possibilities in analyzing SEL trends, teachers can be given more flexibility in how they address SEL (Ashdown & Bernard, 2011; Bailey et al., 2019; Lawson et al., 2019; McKown, 2019; West et al., 2020). Follow-up analysis can help determine if certain techniques teachers employ assists in the presentation of SEL – consistently assessing and contributing to the SEL instruction literature. However, as the results of these models indicate (refer to Tables 11-25), practitioners must stay aware that certain student and rater characteristics may bias the models. Attempts should be made to continue to improve prediction accuracy and reduce these biases by further training and refining the models.

As seen in Appendix D, this study employed a rater error training approach to train the agents in filling out the rating form. In addition to this training form, it would be beneficial to engage in frame-of-reference training with all the agents/raters. This approach involves aligning standards and criterion used by raters when they observe and evaluate performance, helping reduce rating idiosyncrasies. As these models refine over time, they will assist in providing samples of behavioral incidents representing each dimension of growth mindset as well as various levels of performance on those dimensions (i.e., a 2 vs a 4 on the growth mindset developed scale). Some raters may be more lenient than others, which may explain some additional variance in the growth mindset ratings; therefore, controlling for leniency and providing frame-of-reference training should assist model accuracy (Roch et al., 2012).

Limitations

Multiple methodological limitations exist with the current study. The first limitation is the reliance upon agent ratings, who may be prone to biases, and not having an external criterion to compare the growth mindset scores against. Having such a criterion would assist with showing predictive validity and offer more guidance in factors to include in the prediction model. However, as this study was a novel study that conceptualized growth mindset with the Harvard SEL taxonomy (EASEL Lab, 2021b) to be as comprehensive of growth mindset as possible, there was not an idealized external criterion to compare with. A second limitation is the "black box" of the utilized text analytics machine learning algorithms. Many of the computational output throughout the process are not interpretable, or not easily interpretable, by human researchers. Much of the predictive model analysis occurred within this "black box" causing the actual words that resulted in differences between the growth mindset categorizations to not be determined (Miaschi & Dell'Orletta, 2020; Reimers et al, 2019. The *text* package (Kjell et al., 2021) does have functions (e.g., *textProjection* and *textProjectionPlot*) that permit one to begin to understand word differences in order to address this limitation. However, the current study suffered from a high negative skew and range restriction of the growth mindset scale data. The median of the growth mindset score was 16 on a scale of 4 to 20 indicating a negative skew. Only 73 students of 808 scored less than a 10 on the total growth mindset score scale. This skewness in scores may have reduced the reported correlations between the text embeddings and growth mindset scales. This skewness may be explained by the fact this study's data was acquired from an opt-in third-party mentoring program (students in this program may have a higher growth mindset than the general student population) or by agents being lenient raters. Additionally, the skewness may be perpetuated by the inclusion of students who were not engaged or involved in the mentoring program; for example, one student only sent 2 messages in their entire conversation history with their agent/mentor. Future studies may consider a minimum threshold for number of exchanges for inclusion in model analyses. Follow-up studies should acknowledge these limitations in terms of this study's generalizability.

Conclusion

In summary, the results of this study provided proof of the concept that analysis of text message conversations between agents and students can be used to assess a student's level of growth mindset. The results also promote scalability; the predictive models can grow and be refined over time with additional feature engineering, responding to Elias's (2019) concern for the future of SEL. Additionally, while conducting this analysis this research team and the third-party practitioners were able to coordinate reviewing, discussing, and interpreting the data's meaning helping ensure practitioners are obtaining useful and actionable insights responding to some of McKown's (2019) highlighted challenges. Future research into the actual text content that corresponded in differences in growth mindset will assist with practitioners being able to follow-up with necessary interventions to support students. It is suggested that practitioners work with established text analytics researchers for the initial set-up of text analytics analysis to align goals and research rigor; after the initial time and resource lift to set the initial text analytics modeling up, this approach can serve as a feasible low-lift assessment strategy.

References

Almlund, M., Duckworth, A. L., Heckman, J. J., & Kautz, T. D. (2011). Personality psychology and economics. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the Economics of Education* (Vol. 4, pp. 1–181). Elsevier.

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*. https://doi.org/10.48550/arXiv.1904.03323

Al-Maitah, M. (2018). Text analytics for big data using rough-fuzzy soft computing techniques. *Expert Systems 36*(6), e12463. https://doi.org/10.1111/exsy.12463

Ashdown, D. M., & Bernard, M. E. (2011). Can explicit instruction in social and emotional learning skills benefit the social-emotional development, well-being, and academic achievement of young children? *Early Childhood Education Journal 39*, 397-405. https://doi.org/10.1007/s10643-011-0481-x

Bailey, R., Stickle, L., Brion-Meisels, G., & Jones, S. M. (2019). Re-imagining social-emotional learning: Findings from a strategy-based approach. *Phi Delta Kappan 100*(5), 53-58. https://kappanonline.org/re-imagining-social-emotional-learning-findings-from-a-strategy-based-approach/

Brownlee, J. (2017). *What are word embeddings for text*? https://machinelearningmastery.com/what-are-word-embeddings/

Brownlee, J. (2019). *A gentle introduction to the bag-of-words model*. https://machinelearningmastery.com/gentle-introduction-bag-words-model/

Castillo, A. (2021). Natural language processing: Machine learning methods in forensic accounting. *The CPA Journal,* 16-19, https://www.cpajournal.com/2021/08/24/natural-language-processing/

Chen, Y., Perozzi, B., Al-Rfou, R., & Skiena, S. (2013). The expressive power of word embeddings. *arXiv preprint arXiv:1301.3226*. https://doi.org/10.48550/arXiv.1301.3226

Claro, S., Paunesku, D., & Dweck, C. S. (2016). Growth mindset tempers the effects of poverty on academic achievement. *PNAS 113*(31), 8664-8668. https://doi.org/10.1073/pnas.1608207113

Collie, R. J., Shapka, J. D., & Perry, N. E. (2012). School climate and social-emotional learning: Predicting teacher stress, job satisfaction, and teaching efficacy. *Journal of Educational Psychology 104*(4), 1189-1204. http://dx.doi.org/10.1037/a0029356

Deming, D. (2017). The growing importance of social skills in the labor market. Quarterly *Journal of Economics, 132*(4), 1593–1640.

Denham, S. A., & Brown, C. (2010). “Plays nice with others”: Social-Emotional learning and academic success. *Early Education and Development 21*(5), 652-680. https://doi.org/10.1080/10409289.2010.497450

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171-4186. https://doi.org/10.48550/arXiv.1810.04805

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child development, 82*(1), 405-432. https://doi.org/10.1111/j.1467-8624.2010.01564.x

Dweck, C. (2017). *Mindset - updated edition - changing the way you think to fulfil your potential*. New York, NY: Little, Brown Book Group.

EASEL Lab. (2021a). *Identify Related Skills Across Frameworks.* http://exploresel.gse.harvard.edu/compare-terms/

EASEL Lab. (2021b). *Taxonomy Project*. https://easel.gse.harvard.edu/taxonomy-project

Eklund, K., Kilpatrick, K. D., Kilgus, S. P., & Haider, A. (2018). A systematic review of state-level social-emotional learning standards: Implications for practice and research. *School Psychology Review 47*(3), 316-326. https://doi.org/10.17105/SPR-2017.0116.V47-3

Elias, M. J. (2019). What if the doors of every schoolhouse opened to social-emotional learning tomorrow: Reflections on how to feasibly scale up high-quality SEL. *Educational Psychologist 54*(3), 233-245. https://doi.org/10.1080/00461520.2019.1636655

Elliott, S. N., & Gresham, F. M. (2007). *Social Skills Improvement System: Classwide Intervention Program*. Minneapolis, MN: Pearson.

Fang, Q., Nguyen, D., & Oberski, D. L. (2022). Evaluating the Construct Validity of Text Embeddings with Application to Survey Questions. *arXiv preprint arXiv:2202.09166*. https://doi.org/10.48550/arXiv.2202.09166

Ferreira, R., Kovanovic, V., Gasevic, D., & Rolim, V. (2018, June). *Towards combined network and text analytics of student discourse in online discussions*. Paper presented at the 19th International Conference on Artificial Intelligence in Education, London, England. https://doi.org/10.1007/978-3-319-93843-1_9

Hamilton, L. S., Stecher, B. M., Schweig, J., & Baker, G. (2021). *RAND Education Assessment Finder*. RAND Corporation. https://www.rand.org/pubs/tools/TL308.html

Hart, S. C., DiPerna, J. C., Lei, P.-W., & Cheng, W. (2020). Nothing lost, something gained? Impact of a universal social-emotional learning program on future state test performance. *Educational Researcher 49*(1), 5-19. https://doi.org/10.3102%2F0013189X19898721

Harris, Z. S. (1954). Distributional structure. *Word, 10*(2-3), 146-162. https://doi.org/10.1080/00437956.1954.11659520

Heckman, J. J., Humphries, J. E., & Kautz, T. (Eds.) (2014). The myth of achievement tests: GED and the role of character in American life. University of Chicago Press

Heggart, K. (2015). *Developing a growth mindset in teachers and staff*. Edutopia. https://www.edutopia.org/discussion/developing-growth-mindset-teachers-and-staff

Hochanadel, A., & Finamore, D. (2015). Fixed and growth mindset in education and how grit helps students persist in the face of adversity. Journal of International *Education Research 11*(1), 47-50. http://dx.doi.org/10.19030/jier.v11i1.9099

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: applications to nonorthogonal problems. *Technometrics, 12*(1), 69-82. https://doi.org/10.1080/00401706.1970.10488635

Jawahar, G., Sagot, B., & Seddah, D. (2019, July). What does BERT learn about the structure of language?. *In ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*. https://hal.inria.fr/hal-02131630/document

Jia, Q., Cui, J., Xiao, Y., Liu, C., Rashid, P., & Gehringer, E. F. (2021). ALL-IN-ONE: Multi-Task Learning BERT models for Evaluating Peer Assessments. *arXiv preprint arXiv:2110.03895*. https://doi.org/10.48550/arXiv.2110.03895

Jones, S., Bailey, R., Brush, K., & Nelson, B. (2019). *Introduction to the Taxonomy Project: Tools for selecting and aligning SEL frameworks*. https://echidnagiving.org/wp-content/uploads/2020/01/International-Frameworks-for-SEL.pdf

Jones, S. M., Bailey, R., & Jacob, R. (2014). Social-Emotional learning is essential to classroom management. *Phi Delta Kappan 96*(2), 19-24. https://kappanonline.org/social-emotional-learning-essential-classroom-management-jones-bailey-jacobs/

Jones, S. M., & Bouffard, S. M. (2012). Social and emotional learning in schools: From programs to strategies and commentaries. *Social Policy Report 26*(4), 1-33. https://doi.org/10.1002/j.2379-3988.2012.tb00073.x

Kulshrestha, R. (2019). *NLP 101: Word2Vec – Skip-gram and CBOW*. https://towardsdatascience.com/nlp-101-word2vec-skip-gram-and-cbow-93512ee24314

Kim, M.-S., & Park, S. (2020). Growth of fixed mindsets from elementary to middle school: Its relationship with trajectories of academic behavior engagement and academic achievement. *Psychology in the Schools 58*(11), 2175-2188. https://doi.org/10.1002/pits.22583

Kjell, O., Schwartz, H. A., & Giorgi, S. (2021). Text: An R-package for analyzing and visualizing human language using natural language processing and deep learning. *PsyArXiv Preprints*. https://doi.org/10.31234/osf.io/293kt

Lawson, G. M., McKenzie, M. E., Becker, K. D., Selby, L., & Hoover, S. A. (2019). The core components of evidence-based social emotional learning programs. *Prevention Science 20*, 457-467. https://doi.org/10.1007/s11121-018-0953-y

Lexalytics. (2021). *Text Analytics*. https://www.lexalytics.com/technology/text-analytics

Linguamatics. (2021). *What is text mining, text analytics, and natural language processing*? https://www.linguamatics.com/what-text-mining-text-analytics-and-natural-language-processing

Liu, M.-C., & Huang, Y.-M. (2017). The use of data science for education: The case of social-emotional learning. *Smart Learning Environments 4*(1), 1-13. https://doi.org/10.1186/s40561-016-0040-4

McCormick, C., & Ryan, N. (2019). *BERT word embeddings tutorial*. http://mccormickml.com/

McCormick, M. P., Cappella, E., O'Connor, E. E., & McClowry, S. G. (2015). Social-Emotional learning and academic achievement: Using causal methods to explore classroom-level mechanisms. *AERA Open 1*(3), 1-26. https://doi.org/10.1177%2F2332858415603959

McKown, C. (2019). Challenges and opportunities in the applied assessment of student social and emotional learning. *Educational Psychologist 54*(3), 205-221. https://doi.org/10.1080/00461520.2019.1614446

Miaschi, A., & Dell'Orletta, F. (2020, July). Contextual and non-contextual word embeddings: an in-depth linguistic investigation. In *Proceedings of the 5th Workshop on Representation Learning for NLP* (pp. 110-119). https://aclanthology.org/2020.repl4nlp-1.15.pdf

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems, 26*. https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

Mohammed, A. H. K., Jebamikyous, H.-H., Nawara, D., & Kashef, R. (2021). IoT text analytics in smart education and beyond. *Journal of Computing in Higher Education 33*, 779-806. https://doi.org/10.1007/s12528-021-09295-x

Mohler, T. (2020). *The 7 basic functions of text analytics and text mining*. Lexalytics. https://www.lexalytics.com/lexablog/text-analytics-functions-explained

Nasir, M., Dag, A., Young II, W. A., & Delen, D. (2020). Determining optimal skillsets for business managers based on local and global job markets: A text analytics approach. *Decision Sciences Journal of Innovative Education 18*(3), 374-408. https://doi.org/10.1111/dsji.12212

New Jersey School Boards Association. (2017). *The task force on student achievement: advancing education for all children*. https://www.njsba.org/wp-content/uploads/2017/03/Final-Report-March-31.pdf

O'Keefe, P. A., Dweck, C., & Walton, G. (2018, September 10). Having a growth mindset makes it easier to develop new interests. *Harvard Business Review*. Online only. https://hbr.org/2018/09/having-a-growth-mindset-makes-it-easier-to-develop-new-interests

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (EMNLP) (pp. 1532-1543). https://aclanthology.org/D14-1162.pdf

Prodigy. (2021). *Prodigy 101 – everything you need to know.* https://prodi.gy/docs

Pustejovsky, J., & Stubbs, A. (2012). The Basics. In *Natural Language Annotation for Machine Learning. O'Reilly*. (pp. 1-33). O'Reilly. https://www.oreilly.com/library/view/natural-language-annotation/9781449332693/ch01.html

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Reimers, N., Schiller, B., Beck, T., Daxenberger, J., Stab, C., & Gurevych, I. (2019). Classification and clustering of arguments with contextualized word embeddings. *arXiv preprint arXiv:1906.09821*. https://doi.org/10.48550/arXiv.1906.09821

Roch, S. G., Woehr, D. J., Mishra, V., & Kieszczynska, U. (2012). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology, 85*(2), 370-395. https://doi.org/10.1111/j.2044-8325.2011.02045.x

Student Success Agency. (2021). Agents. https://studentsuccess.co/agents/#agent-demographics

Titenok, Y. (2022). *Natural language processing vs text mining*. https://sloboda-studio.com/blog/natural-language-processing-vs-text-mining/

Valcarcel, C. V., Holmes, J., Berliner, D. C., & Koerner, M. (2021). The value of student feedback in open forums: A natural language analysis of descriptions of poorly rated teachers. *Education Policy Analysis Archives 29*(79), https://doi.org/10.14507/epaa.29.6289

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods, 45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

Wiedemann, G., Remus, S., Chawla, A., & Biemann, C. (2019). Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. *arXiv preprint arXiv:1909.10430*. https://doi.org/10.48550/arXiv.1909.10430

West, M. R., Pier, L., Fricke, H., Hough, H., Loeb, S., Meyer, R. H., & Rice, A. B. (2020). Trends in student social-emotional learning: Evidence from the first large-scale panel student survey. *Education Evaluation and Policy Analysis 42*(2), 279-303. https://doi.org/10.3102%2F0162373720912236

Yang, C., Bear, G. G., & May, H. (2018). Multilevel associations between school-wide social-emotional learning approach and student engagement across elementary, middle, and high schools. *School Psychology Review 47*(1), 45-61. https://doi.org/10.17105/SPR-2017-0003.V47-1

Appendix A

IRB Approval

Appalachian STATE UNIVERSITY

INSTITUTIONAL REVIEW BOARD
Office of Research Protections
ASU Box 32068
Boone, NC 28608
828.262.2692
Web site: http://researchprotections.appstate.edu
Email: irb@appstate.edu
Federalwide Assurance (FWA) #00001076

To: Adam Blackburn
Psychology ,
CAMPUS EMAIL

From: IRB Administration
Date: 2/04/2022
RE: Determination that Research or Research-Like Activity does not require IRB Approval
Agrants #:
Grant Title:

STUDY #: 22-0154
STUDY TITLE: Reading Between the Lines: A Text Analytics Exploration of Social Emotional Learning

The IRB determined that the activity described in the study materials does not constitute human subject research as defined by University policy and the federal regulations [45 CFR 46.102 (e or l)] and does not require IRB approval.

Study Regulatory and other findings:

Study determined to be NHSR due to receipt of de-identified information. Please see administrative attachments for more information about data transmission.

This determination may no longer apply if the activity changes. IRB approval must be sought and obtained for any research with human participants.

If you have any questions about this determination, please contact the IRB Administration at 828-262-2692 or irb@appstate.edu.

Thank you.

Appendix B

Tables

Table 1

Harvard SEL's Domain Subdomains

Cognitive	• Attention Control • Working Memory and Planning Skills • Inhibitory Control • Cognitive Flexibility • Critical Thinking
Social	• Understanding Social Cues • Conflict Resolution/Social Problem Solving • Prosocial/Cooperative Behavior
Perspectives	• Optimism • Gratitude • Openness • Enthusiasm/Zest
Emotion	• Emotional Knowledge and Expression • Emotional and Behavioral Regulation • Empathy/Perspective Taking
Values	• Ethical Values • Performance Values • Civic Values • Intellectual Values
Identity	• Self-Knowledge • Purpose • Self-Efficacy/Growth Mindset • Self-Esteem

Table 2

Harvard SEL's Identity Subdomains and Competencies

Self-Knowledge	• Identifies and understands personality/character traits • Recognizes and understands one's own strengths and weaknesses • Honest about what you know and don't know • Develop and maintain a coherent sense of self and roles over time • Identifies and understands one's interests and preferences
Purpose	• Considers existential questions (e.g., what is the purpose of my life, what is my life passion, what is happiness, what is my place in the world, etc.) • Imagines the future; formulates life goals and way to pursue them • Expresses and derives comfort from a belief in something greater than self
Self-Efficacy/ Growth Mindset	• Believes that intellectual abilities and personality traits are qualities that can be developed and improved • Expresses confidence in oneself and one's ability to improve or succeed • Sees challenges as things that one can take on and overcome with time and effort • Belief that one has a choice (agency)
Self-Esteem	• Feels a sense of belonging; feels valued by others in the community • Extends kindness and understanding to oneself (e.g., has self-compassion, emotional self-respect, etc.) • Forgives oneself for errors and mistakes (e.g., accepts and moves on from past actions) • Demonstrates physical self-respect by maintaining good hygiene • Understands the effects of risks behaviors (e.g., drugs, alcohol, tobacco, sex, etc.) on their body and use that information to make responsible choices • Believes that one is not defined by one's thoughts, emotions, or circumstances

Table 3

Steps to Prepare a Source for Text Analytics and NLP

Step Number	Step Name	Step Description
1	Language Identification	Identifying the text is written in English
2	Tokenization	Breaking a sentence into parts/tokens. For example, separating at each individual word, pulling out punctuation, identifying hyperlinks
3	Sentence Breaking	Analyzing punctuation to determine when sentences end (e.g., does the period at end of Dr. end the sentence?)
4	Part of Speech Tagging	Tagging each token with its corresponding part of speech (e.g., noun, verb, preposition)
5	Chunking	Assigning part of speech tagged tokens to phrases (i.e., piecing together tokens into noun phrases, verb phrases_
6	Syntax Parsing	Determining the structure of the sentence and help prepare sentiment analysis (e.g., Apple was doing poorly until Steve Job... - “Apple” is negative while “Steve Jobs” is neutral; Because Apple was doing poorly, Steve Jobs... - “Apple” is negative while “Steve Jobs” is neutral)
7	Sentence Chaining	Connecting related sentences together based on strength of association to an overall topic

Table 4

Student Demographics and Frequency Tables

Demographic Characteristic	*n*	Missing	Mean	Median	SD	Min	Max	Percentiles 25th	Percentiles 50th	Percentiles 75th
Student Gender	808	0								
Student State	808	0								
Student Age	795	13	16.80	17.00	1.48	13.00	22.00	16.00	17.00	18.00
Student Time in Program	808	0	530.00	474.00	302.08	84.00	1566.00	316.80	473.50	792.00

Note. Student Time in Program is measured in days.

Frequencies of Student Gender

Levels	*Counts*	*% of Total*
Female	480	59.40%
Male	328	40.60%

Frequencies of Student State

Levels	*Counts*	*% of Total*
Alabama	73	9.00%
Arizona	91	11.30%
California	25	3.10%
Florida	35	4.30%
Georgia	227	28.10%
Illinois	4	0.50%
Kentucky	10	1.20%
Louisiana	3	0.40%
Montana	37	4.60%
Nevada	98	12.10%
North Carolina	27	3.30%
Oklahoma	41	5.10%
Pennsylvania	1	0.10%
Texas	112	13.90%
Virginia	1	0.10%
Washington	23	2.80%

Table 5

Agent Demographics and Frequency Tables

Demographic Characteristic	*n*	Missing	Mean	Median	SD	Min	Max	Percentiles 25th	50th	75th
Agent Gender	24	0								
Agent Ethnicity	24	0								
Agent Age	24	0	23.50	23.00	2.25	21.00	32.00	22.00	23.00	24.30
Agent Time in Program	24	0	914.60	899.00	372.49	434.00	1629.00	483.30	899.00	1264.00
Number of Students	808	0	44.80	32.00	25.20	14.00	83.00	25.00	32.00	80.00

Note. Agent Time in Program is measured in days.

Frequencies of Agent Gender

Levels	*Counts*	*% of Total*
Female	15	62.50%
Male	9	37.50%

Frequencies of Agent Ethnicity

Levels	*Counts*	*% of Total*
African American	4	16.70%
Arab	1	4.20%
Asian	5	20.80%
Caucasian	7	29.20%
Eastern European	1	4.20%
Hispanic	2	8.30%
Indian	2	8.30%
Western European	2	8.30%

Table 6

Text Message Conversation Descriptive Statistics

Demographic Characteristic	*n*	Missing	Mean	Median	SD	Min	Max	Percentiles		
								25th	50th	75th
Total Messages	808	0	433.50	285.00	555.28	13.00	10259.00	148.75	285.00	540.00
Student-Initiated Messages	808	0	195.29	115.00	318.92	2.00	6793.00	56.75	115.00	235.25
Agent-Initiated Messages	808	0	238.21	167.00	250.52	11.00	3466.00	85.75	167.00	298.25
Avg Texts sent by Student per Day	808	0	.85	.64	1.05	.04	20.44	.40	.64	1.03
Student-Initiated Percent of Total	808	0	.41	.41	.08	.07	.80	.37	.41	.46
Agent-Initiated Percent of Total	808	0	.59	.59	.08	.20	.93	.54	.59	.63

Table 7

Agent Growth Mindset Rating Score Distributions

Description	*n*	*Mean*	*Median*	*SD*	*Variance*	*IQR*	*Range*	*Min*	*Max*
GM Developed	808	3.92	4.00	1.05	1.10	2.00	4.00	1.00	5.00
GM Self-Efficacy	808	3.88	4.00	1.09	1.18	2.00	4.00	1.00	5.00
GM Time and Effort	808	3.96	4.00	1.05	1.11	2.00	4.00	1.00	5.00
GM Independent Choice	808	3.92	4.00	1.09	1.19	2.00	4.00	1.00	5.00
GM Total Score	808	15.69	16.00	3.99	15.91	7.00	16.00	4.00	20.00

Table 8

Correlation matrix of Growth Mindset Scores

Variable	*n*	Mean	SD	12	13	14	15	16
12. GM Developed	808	3.92	1.05	--				
13. GM Self-Efficacy	808	2.88	1.09	.82***	--			
14. GM Time and Effort	808	2.96	1.05	.84***	.83***	--		
15. GM Independent Choice	808	2.92	1.09	.81***	.81***	.84***	--	
16. GM Total	808	15.69	3.99	.93***	.93***	.94***	.93***	--

Note. Correlations with GM Total are spuriously inflated as they are part-whole correlations. *** $p < .001$

Table 9

Correlation matrix of Student/Agent Characteristics and Text Message Characteristics

Variable	*n*	Mean	SD	1	2	3	4	5	6	7	8	9	10	11
1. Student Age	795	16.76	1.48	--										
2. Student Time in Program	808	530.02	302.08	.25***	--									
3. Agent Age	808	23.57	1.98	-.03	-.05	--								
4. Agent Time in Program	808	1029.62	279.76	-.003	.27***	-.10**	--							
5. Number of Students	808	44.84	25.18	.002	.13***	-.05	.62***	--						
6. Total Messages	808	433.50	555.28	.12***	.40***	-.07*	.09*	.07	--					
7. Student-Initiated Messages	808	195.29	318.92	.09*	.30***	-.05	.05	.03	.98***	--				
8. Agent-Initiated Messages	808	238.21	250.52	.15***	.49***	-.09**	.136***	.11**	.97***	.90***	--			
9. Avg Texts sent by Student per Day	808	.39	.61	.001	-.05	-.07	-.03	.01	.70***	.74***	.61***	--		
10. Student-Initiated Percent of Total	808	.41	.08	.03	-.03	-.02	-.05	-.13	.37***	.40***	.32***	.46***	--	
11. Agent-Initiated Percent of Total	808	.59	.08	-.03	.03	.02	.05	13***	-.37***	-.40***	-.32***	-.46***	-1.00	--

Note. Correlations between the text message characteristics (6, 7, 8, 10, and 11) are spuriously inflated as they are part-whole correlations and compare some redundant information * $p < .05$, ** $p < .01$, *** $p < .001$

Table 10

Correlation matrix of Student/Agent/Text Message Characteristics and Growth Mindset Scores

Variable	12. GM Developed	13. GM Self-Efficacy	14. GM Time and Effort	15. GM Independent Choice	16. GM Total
1. Student Age	.15***	.18***	.18***	.17***	.18***
2. Student Time in Program	0.03	.08*	.08*	.08*	.07*
3. Agent Age	0.002	-0.04	-0.03	-0.02	-0.02
4. Agent Time in Program	0.05	.14***	.09*	.09**	.10**
5. Number of Students	-0.02	-0.02	-0.04	-0.02	-0.03
6. Total Messages	0.008	0.01	0.01	0.05	0.02
7. Student-Initiated Messages	-0.01	-0.01	-0.01	0.03	<0.01
8. Agent-Initiated Messages	0.02	0.03	0.03	0.06	0.04
9. Avg Texts sent by Student per Day	-0.03	-0.04	-0.04	-0.03	-0.04
10. Student-Initiated Percent of Total	0.20***	0.19***	0.19***	0.18***	0.20***
11. Agent-Initiated Percent of Total	-0.20***	-0.19***	-.019***	-0.18***	-0.20***

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

Table 11

Word Embedding Prediction Model Fit Indices

GMS Outcome Variable	R	df	*p*	*t*	p-holm	RMSE	Normalized RMSE
GM Developed	.40	806	< .001	12.465	< .001	0.96	.24
GM Self-Efficacy	.37	806	< .001	11.301	< .001	1.01	.25
GM Time and Effort	.41	806	< .001	12.812	< .001	0.96	.24
GM Independent Choice	.40	806	< .001	12.234	< .001	1.00	.25
GM Total Score	.43	806	< .001	13.464	< .001	3.60	.23

Table 12

Growth Mindset Developed Scale Bias Assessment One-Way ANOVAs

Model		Sum of Squares	df	Mean Square	F	*p*	η^2
GM Developed x Student Gender							
	Student Gender	1.33	1	1.33	1.44	.231	.002
	Residuals	745.94	806	.93			
GM Developed x Agent Gender							
	Agent Gender	1.04	1	1.04	1.12	.280	.001
	Residuals	746.24	806	.93			
GM Developed x Agent Ethnicity							
	Agent Ethnicity	29.90	7	4.28	4.77	<.001	.04
	Residuals	717.30	800	.90			

Table 13

GM Developed x Agent Ethnicity Estimated Marginal Means Table

			95% Confidence Interval	
Agent Ethnicity	Mean	SE	Lower	Upper
African American	.28	.09	.10	.46
Arab	.15	.23	-.30	.60
Asian	-.02	.06	-.14	.10
Caucasian	-.28	.07	-.41	-.14
Eastern European	.23	.11	.02	.44
Hispanic	-.10	.14	-37	.17
Indian	.14	.10	-.06	.34
Western European	-.05	.15	-.34	.24

Table 14

Growth Mindset Developed Scale Bias Assessment Correlation Matrix

Variable	GM Developed Residuals
1. Student Age	0.08*
2. Student Time in Program	0.04
3. Agent Age	-0.02
4. Agent Time in Program	0.05
5. Number of Students	0.04
6. Total Messages	0.04
7. Student-Initiated Messages	0.03
8. Agent-Initiated Messages	0.06
9. Avg Texts sent by Student per Day	0.003
10. Student-Initiated Percent of Total	0.18***
11. Agent-Initiated Percent of Total	-0.18***

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

Table 15

Growth Mindset Self-Efficacy Scale Bias Assessment One-Way ANOVAs

Model		Sum of Squares	df	Mean Square	F	*p*	η^2
GM SE x Student Gender							
	Student Gender	0.54	1	0.54	.53	.469	.001
	Residuals	822.73	806	1.02			
GM SE x Agent Gender							
	Agent Gender	0.51	1	.051	.50	.482	.001
	Residuals	822.76	806	1.02			
GM SE x Agent Ethnicity							
	Agent Ethnicity	24.4	7	3.48	4.49	.001	.030
	Residuals	798.9	800	1.00			

Table 16

GM SE x Agent Ethnicity Estimated Marginal Means Table

			95% Confidence Interval	
Agent Ethnicity	Mean	SE	Lower	Upper
African American	.18	.10	-.01	.37
Arab	.48	.24	.005	.96
Asian	-.06	.07	-.19	.07
Caucasian	-.16	.07	-.30	-.02
Eastern European	.11	.11	-.11	.32
Hispanic	.09	.14	-.19	.38
Indian	.21	.11	-.006	.42
Western European	-.40	.16	-.70	-.09

Table 17

Growth Mindset Self-Efficacy Scale Bias Assessment Correlation Matrix

Variable	GM SE Residuals
1. Student Age	0.12**
2. Student Time in Program	0.09**
3. Agent Age	-0.05
4. Agent Time in Program	0.13***
5. Number of Students	0.02
6. Total Messages	0.05
7. Student-Initiated Messages	0.02
8. Agent-Initiated Messages	0.07
9. Avg Texts sent by Student per Day	-0.005
10. Student-Initiated Percent of Total	0.18***
11. Agent-Initiated Percent of Total	-0.18***

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

Table 18

Growth Mindset Time and Effort Scale Bias Assessment One-Way ANOVAs

Model		Sum of Squares	df	Mean Square	F	*p*	η^2
GM Time Effort x Student Gender							
	Student Gender	.003	1	.003	.003	.959	<.001
	Residuals	743	806	.92			
GM Time Effort x Agent Gender							
	Agent Gender	.003	1	.003	.003	.953	<.001
	Residuals	743	806	1.02			
GM Time Effort x Agent Ethnicity							
	Agent Ethnicity	40.9	7	5.85	6.67	<.001	.06
	Residuals	702.1	800	.88			

Table 19

GM Time Effort x Agent Ethnicity Estimated Marginal Means Table

			95% Confidence Interval	
Agent Ethnicity	Mean	SE	Lower	Upper
African American	.35	.09	.17	.53
Arab	.23	.23	-.22	.67
Asian	-.12	.06	-.24	.0006
Caucasian	-.27	.07	-.40	-.14
Eastern European	.13	.10	-.07	.34
Hispanic	.21	.14	-.06	.48
Indian	.23	.10	.03	.43
Western European	-.23	.15	-.51	.06

Table 20

Growth Mindset Time and Effort Scale Bias Assessment Correlation Matrix

Variable	GM TE Residuals
1. Student Age	.08*
2. Student Time in Program	.08*
3. Agent Age	-0.04
4. Agent Time in Program	.08*
5. Number of Students	-0.005
6. Total Messages	0.04
7. Student-Initiated Messages	0.02
8. Agent-Initiated Messages	0.07
9. Avg Texts sent by Student per Day	-0.02
10. Student-Initiated Percent of Total	.16***
11. Agent-Initiated Percent of Total	-.16***

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

Table 21

Growth Mindset Independent Choice Scale Bias Assessment One-Way ANOVAs

Model		Sum of Squares	df	Mean Square	F	*p*	η^2
GM IC x Student Gender							
	Student Gender	.004	1	.004	.004	.95	<.001
	Residuals	813.9	806	1.01			
GM IC x Agent Gender							
	Agent Gender	.01	1	.01	.01	.92	<.001
	Residuals	813.9	806	1.01			
GM IC x Agent Ethnicity							
	Agent Ethnicity	38.4	7	5.49	5.67	<.001	.05
	Residuals	775.5	800	.97			

Table 22

GM IC x Agent Ethnicity Estimated Marginal Means Table

			95% Confidence Interval	
Agent Ethnicity	Mean	SE	Lower	Upper
African American	.25	.10	.07	.44
Arab	.43	.24	-.04	.90
Asian	-.07	.06	.19	.06
Caucasian	-.29	.07	-.43	-.15
Eastern European	.16	.11	-.05	.38
Hispanic	.12	.14	-.16	.40
Indian	.24	.11	.03	.45
Western European	-.30	.15	-.60	.004

Table 23

Growth Mindset Independent Choice Scale Bias Assessment Correlation Matrix

Variable	GM IC Residuals
1. Student Age	.10**
2. Student Time in Program	.07*
3. Agent Age	-0.04
4. Agent Time in Program	.07*
5. Number of Students	0.009
6. Total Messages	0.07
7. Student-Initiated Messages	0.05
8. Agent-Initiated Messages	.09**
9. Avg Texts sent by Student per Day	-0.004
10. Student-Initiated Percent of Total	.17***
11. Agent-Initiated Percent of Total	-.17***

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

Table 24

Growth Mindset Total Scale Bias Assessment One-Way ANOVAs

Model		Sum of Squares	df	Mean Square	F	*p*	η^2
GM Total x Student Gender							
	Student Gender	3.98	1	3.98	1.24	.266	.002
	Residuals	2583.55	806	3.21			
GM Total x Agent Gender							
	Agent Gender	3.36	1	3.36	.26	.611	<.001
	Residuals	10489.67	806	13.01			
GM Total x Agent Ethnicity							
	Agent Ethnicity	519	7	74.1	5.94	<.001	.05
	Residuals	9974	800	12.5			

Table 25

GM Total x Agent Ethnicity Estimated Marginal Means Table

			95% Confidence Interval	
Agent Ethnicity	Mean	SE	Lower	Upper
African American	1.09	.35	.42	1.77
Arab	1.38	.86	-.30	3.06
Asian	-.28	.23	-.73	.17
Caucasian	-1.05	.25	-1.54	-.55
Eastern European	.62	.40	-.16	1.39
Hispanic	.33	.51	-.67	1.33
Indian	.80	.38	.05	1.55
Western European	-1.01	.55	-2.09	.07

Table 26

Growth Mindset Total Scale Bias Assessment Correlation Matrix

Variable	GM Total Residuals
1. Student Age	.09*
2. Student Time in Program	.08*
3. Agent Age	-0.04
4. Agent Time in Program	.09*
5. Number of Students	0.02
6. Total Messages	0.05
7. Student-Initiated Messages	0.03
8. Agent-Initiated Messages	.08*
9. Avg Texts sent by Student per Day	-0.01
10. Student-Initiated Percent of Total	.18***
11. Agent-Initiated Percent of Total	-.18***

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

Appendix C

Figures

Figure 1

Depiction of SEL Rating Form Provided to Agents

			Growth Mindset: "In a growth mindset, people believe that their most basic abilities can be developed through dedication and hard work – brains and talent are just starting points." (Dweck, 2017)							
Student Name	Agent Name	Rating Scale: Please rate the following competencies from 1-5: 1) Strongly Disagree 2) Disagree 3) Neither Agree Nor Disagree 4) Agree 5) Strongly Agree	This student believes that intellectual abilities and qualities can be developed and improved, rather than a skillset they are naturally stuck with	This student exhibits self-efficacy, confidence they have the ability to encounter, identify, and learn to overcome challenges they face throughout life.	This student believes that with enough time and effort, they can overcome most challenges they come across.	This student believes that they have the ability to make independent choices that influence their current and future life (i.e., this student exhibits an internal locus of control rather than external).	Growth Mindset Score	Please provide rationale for your ratings for each competency and overall growth mindset score.	How confident are you in rating Growth Mindset for this Student? (1-5) Please rate your confidence in rating Growth Mindset for this student from 1-5: 1) Very Unconfident 2) Unconfident 3) Neither Confident Nor Unconfident 4) Confident 5) Very Confident	Explain your confidence level with this student

Figure 2

Growth Mindset Developed Subscale Histogram

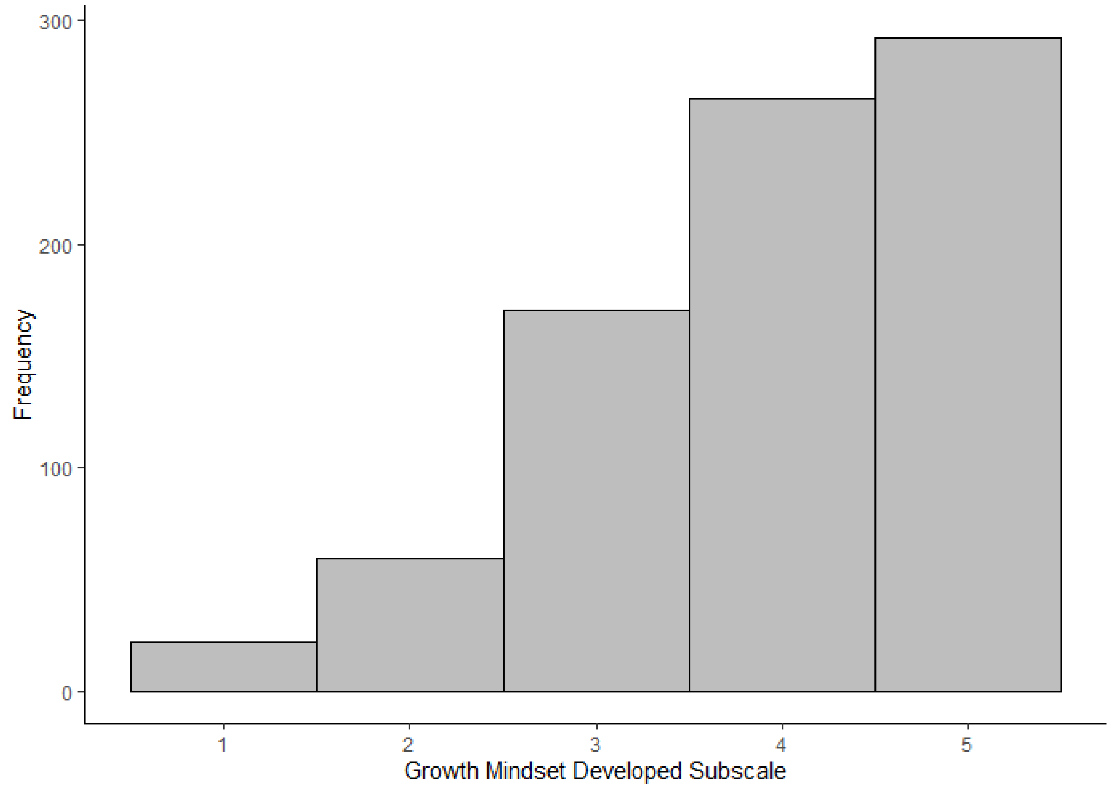


Figure 3

Growth Mindset Self Efficacy Subscale Histogram

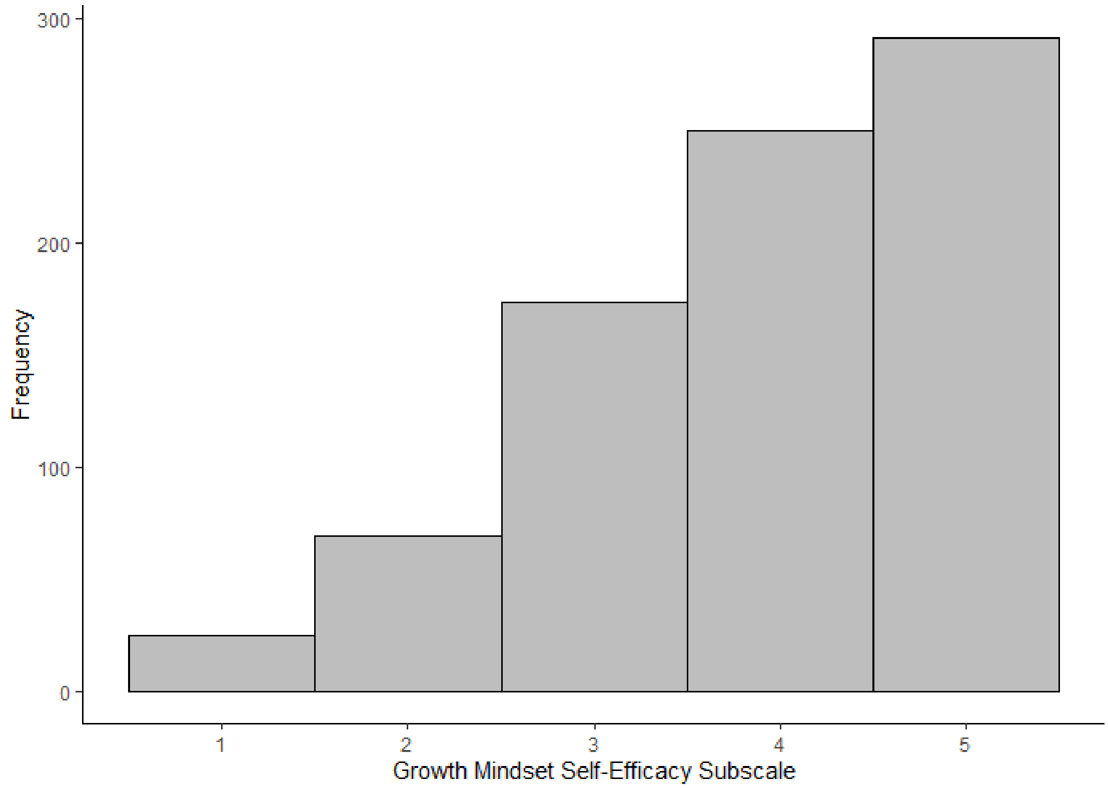


Figure 4

Growth Mindset Time and Effort Subscale Histogram

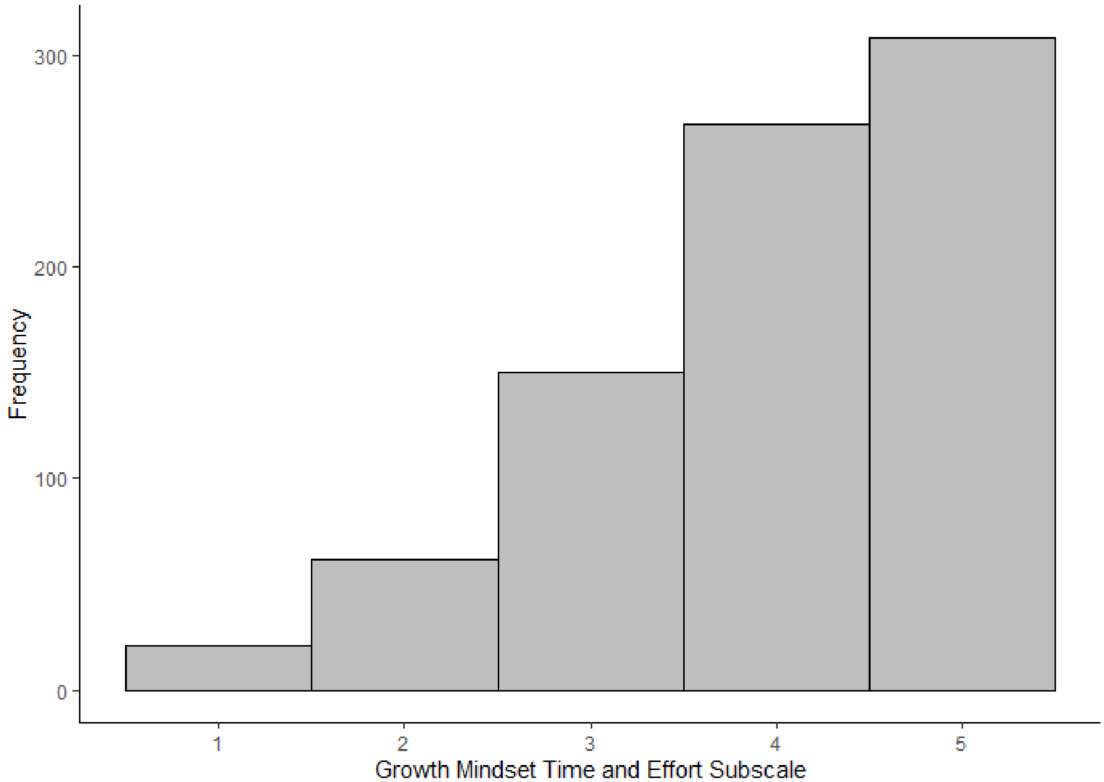


Figure 5

Growth Mindset Independent Choice Subscale Histogram

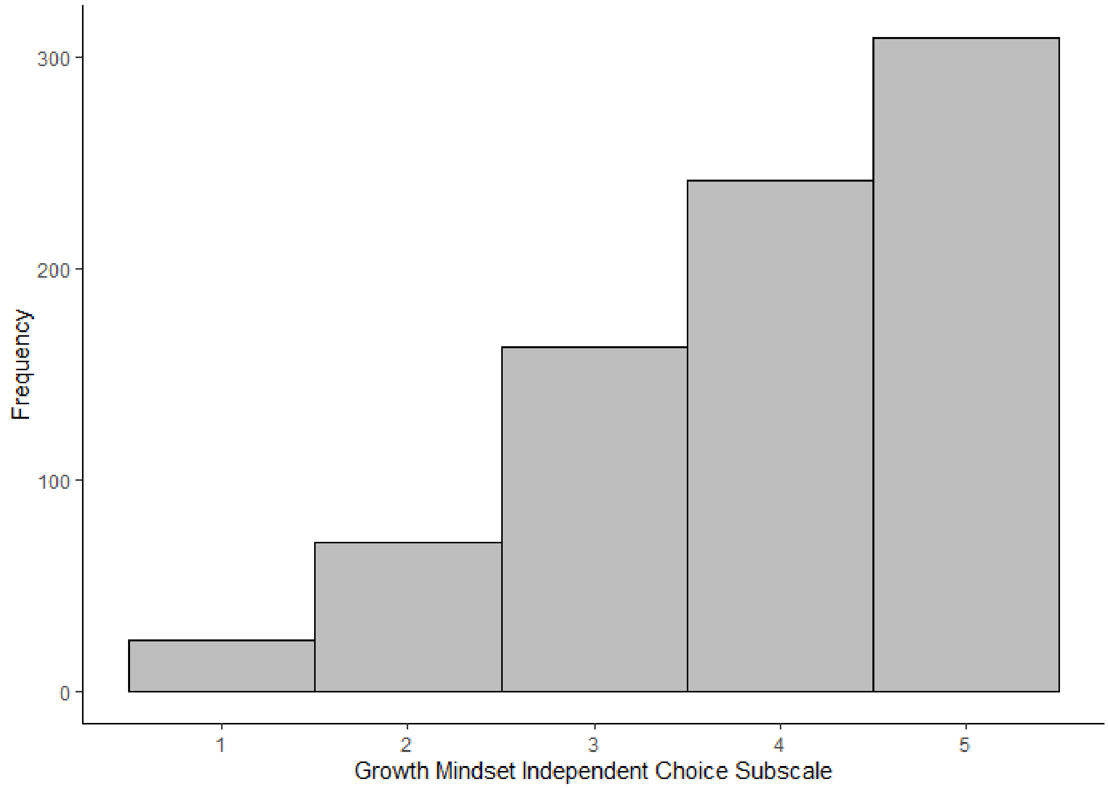


Figure 6

Growth Mindset Total Score Histogram

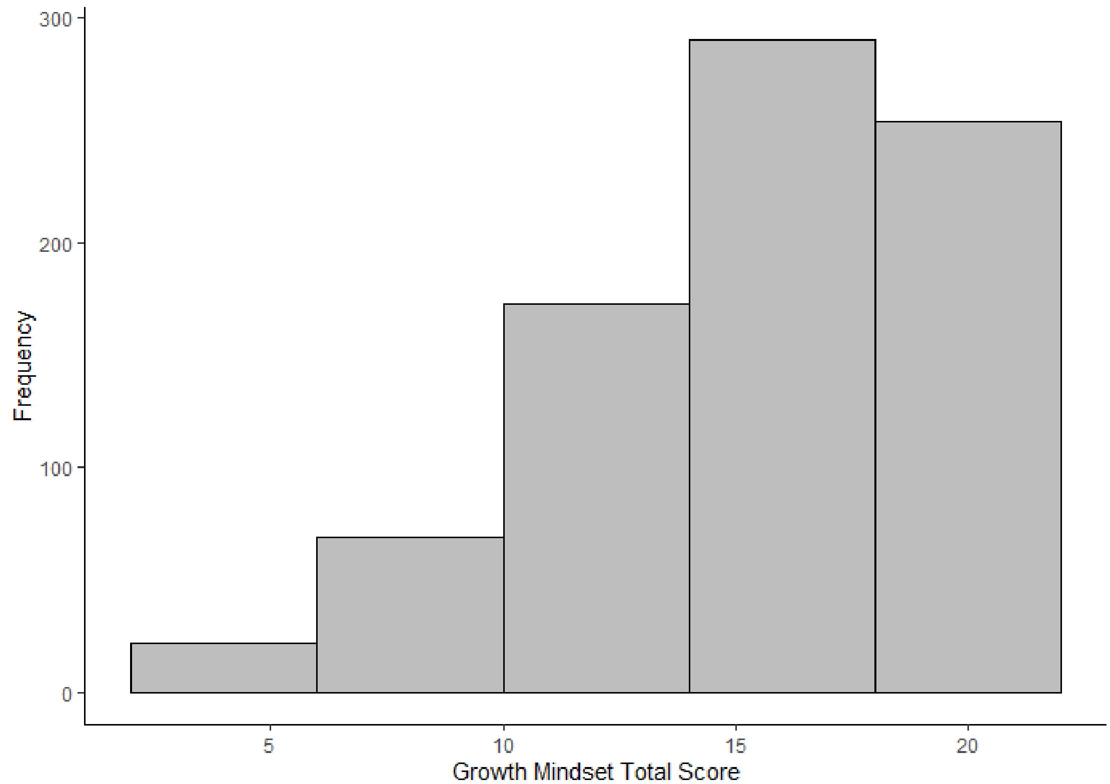


Figure 7

Growth Mindset Developed Prediction Residuals Histogram

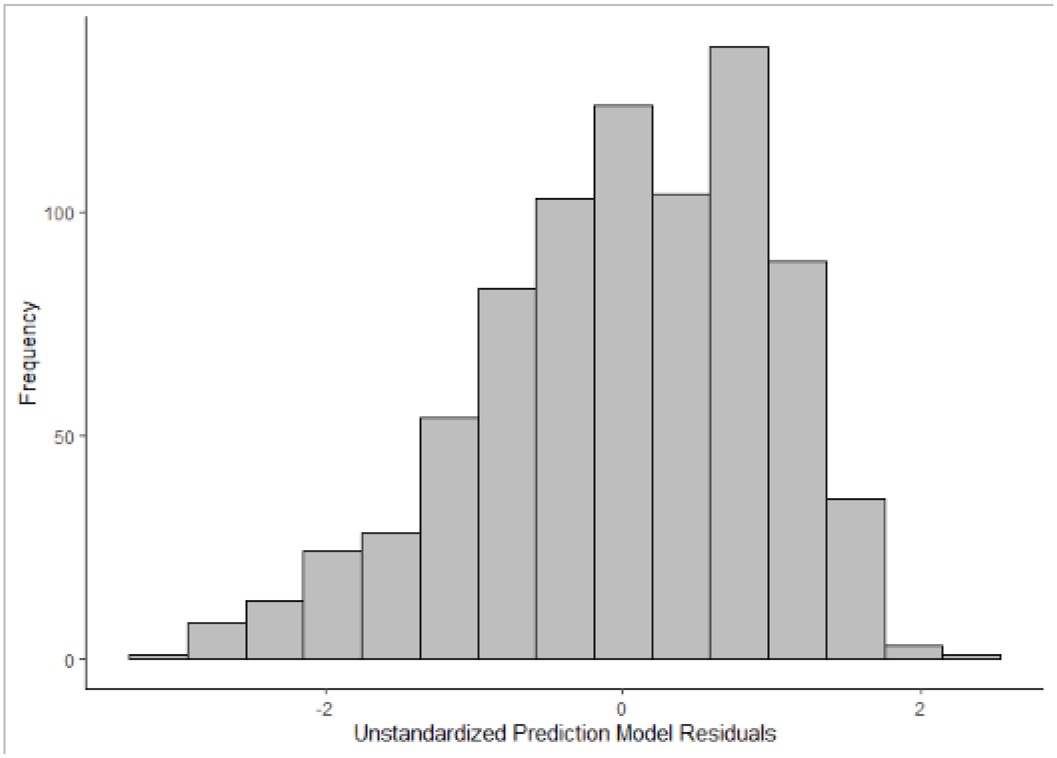


Figure 8

Growth Mindset Developed Residual by Fitted Value Scatterplot

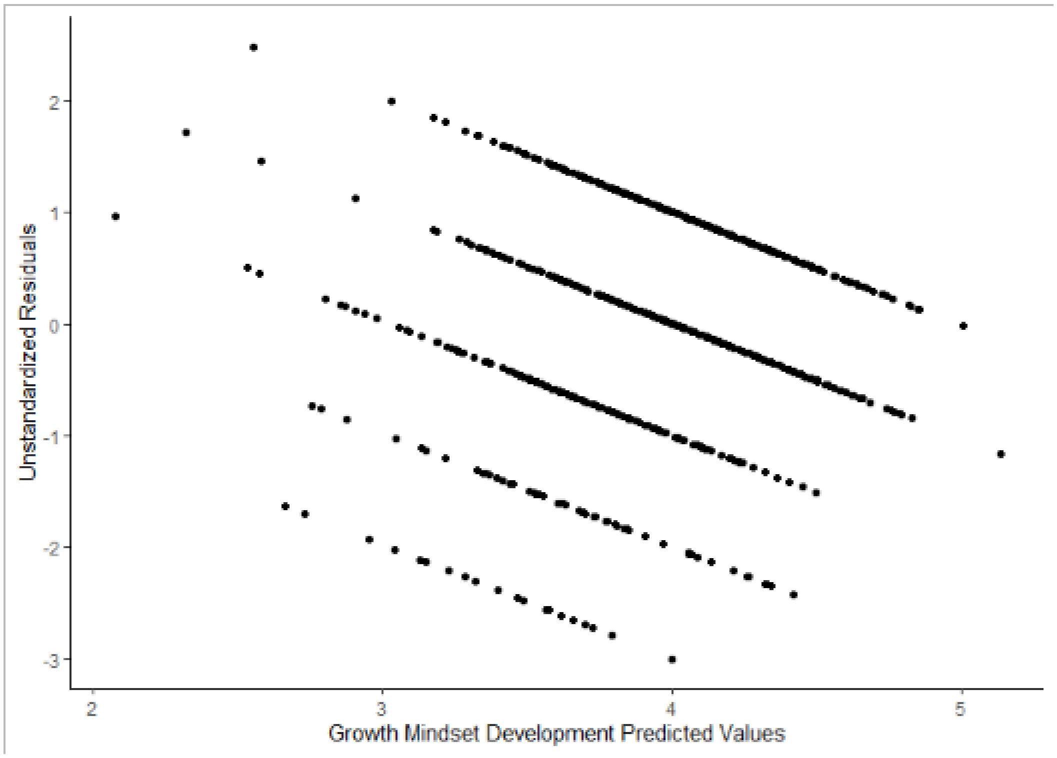


Figure 9

Growth Mindset Self-Efficacy Prediction Residuals Histogram

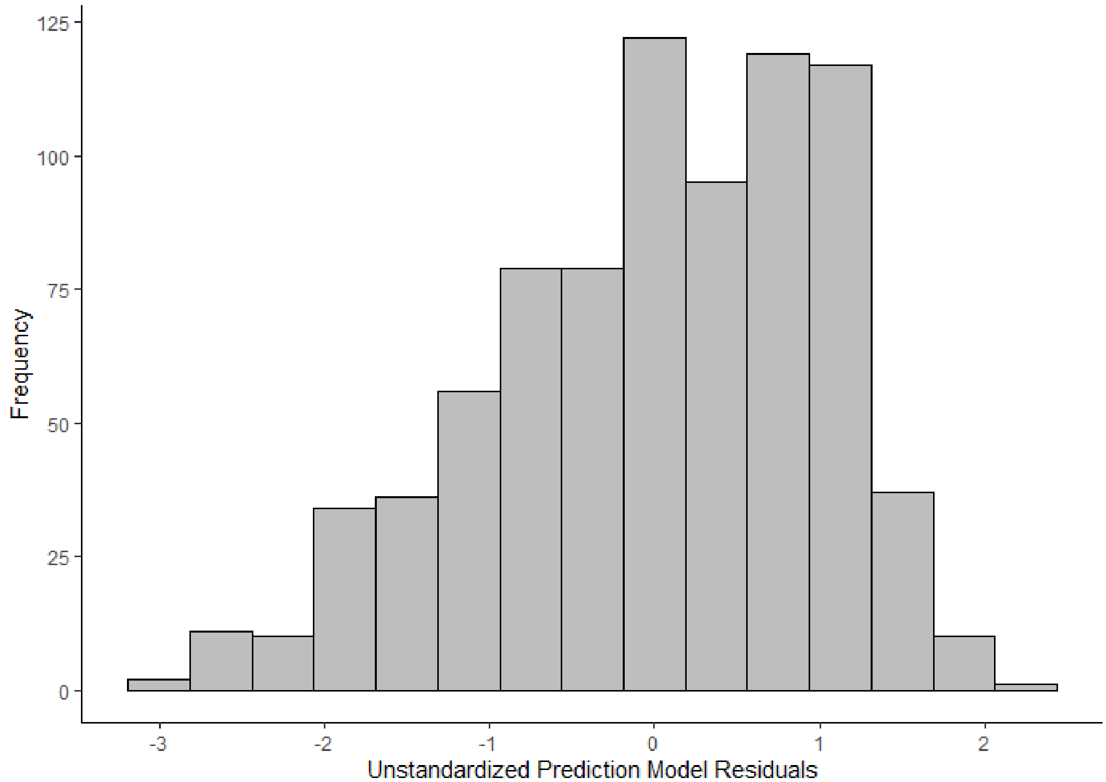


Figure 10

Growth Mindset Self-Efficacy Residual by Fitted Value Scatterplot

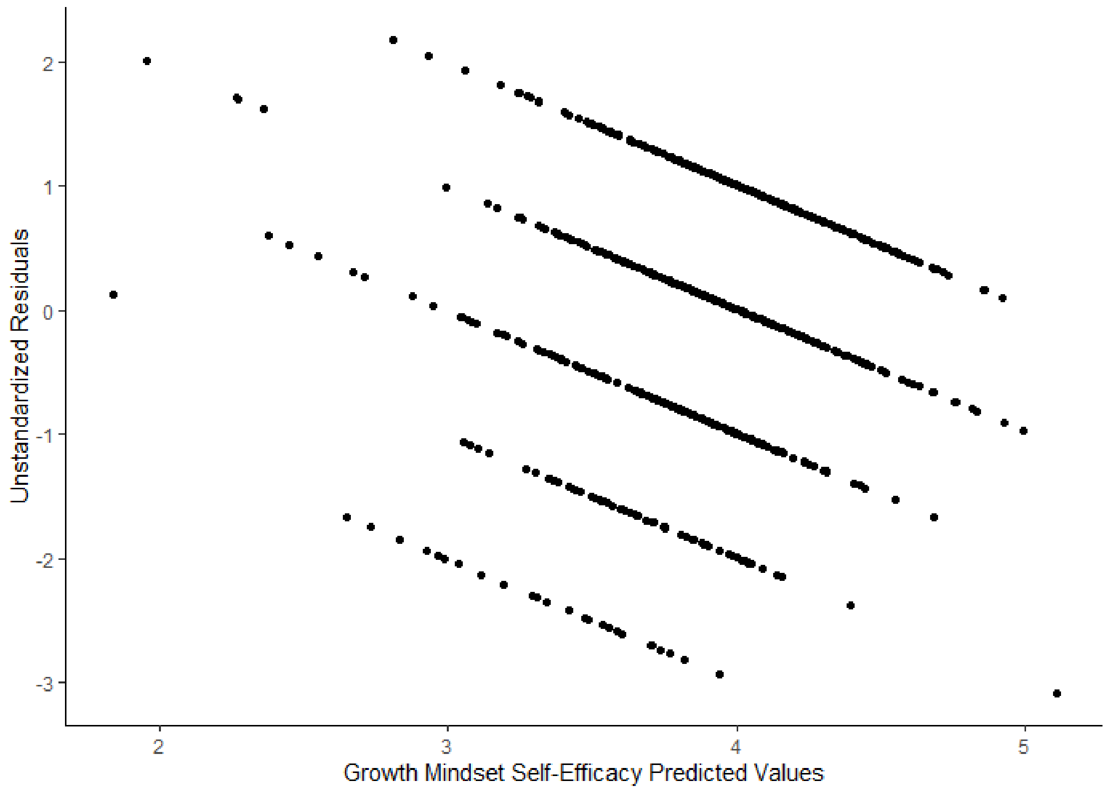


Figure 11

Growth Mindset Time and Effort Prediction Residuals Histogram

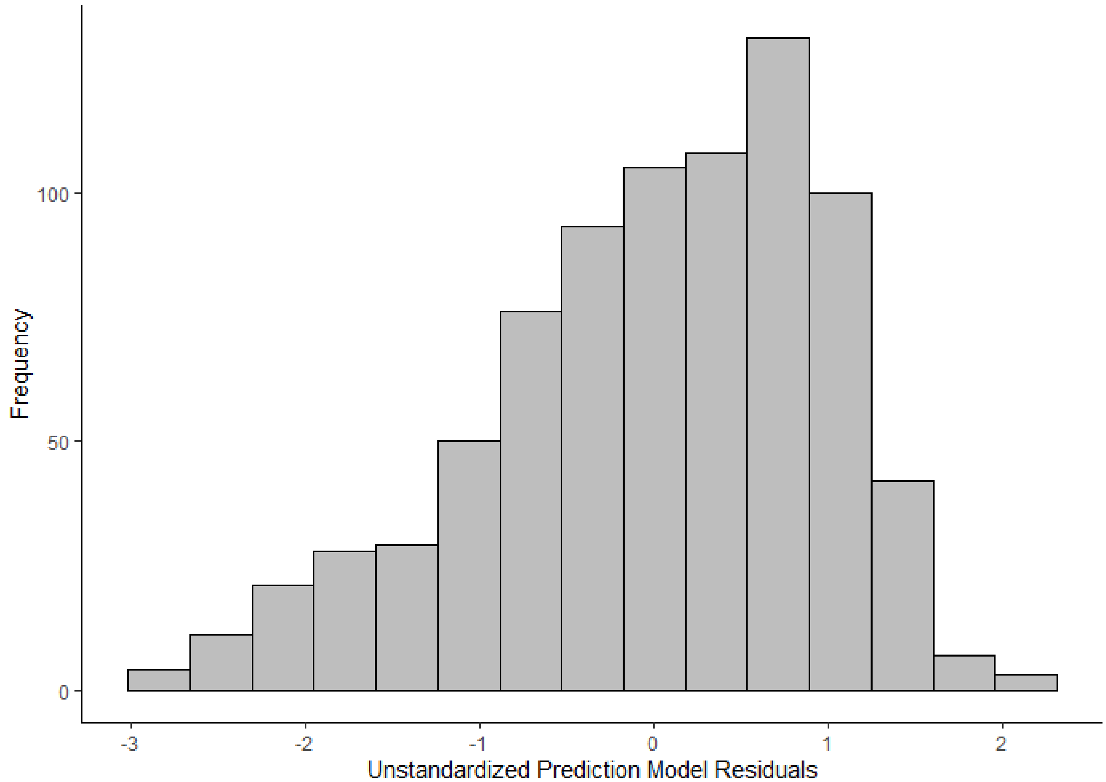


Figure 12

Growth Mindset Time and Effort Residual by Fitted Value Scatterplot

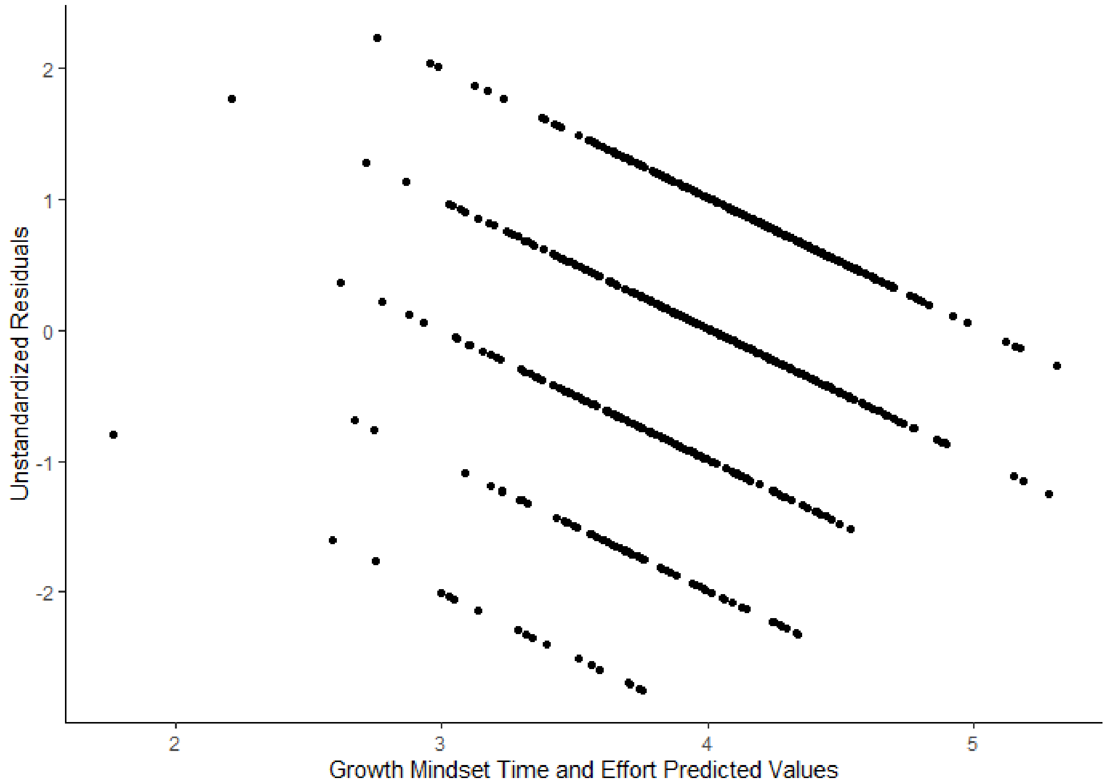


Figure 13

Growth Mindset Independent Choice Prediction Residuals Histogram

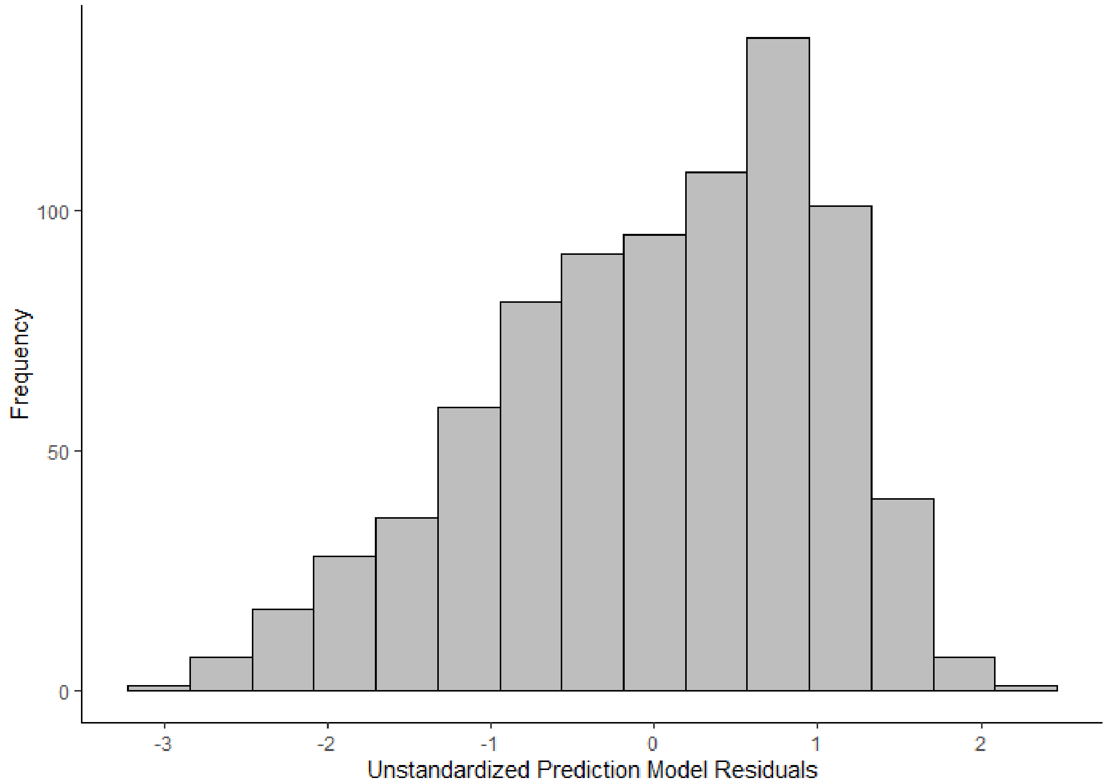


Figure 14

Growth Mindset Independent Choice Residual by Fitted Value Scatterplot

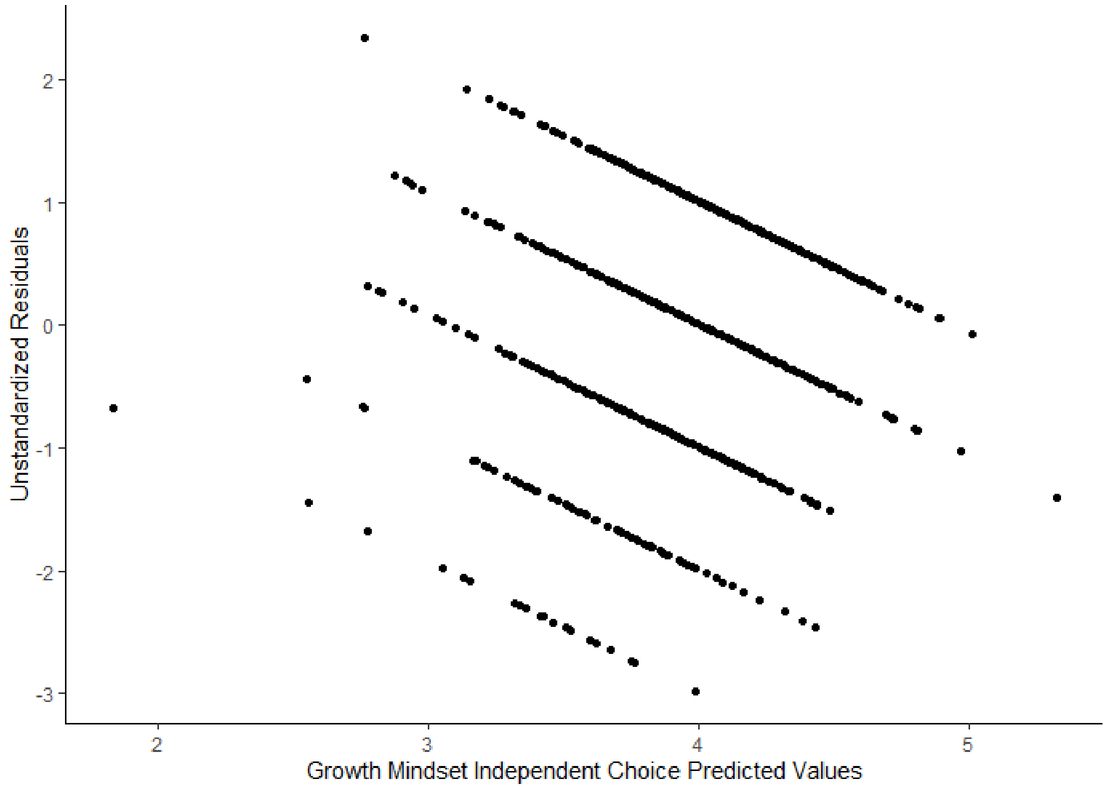


Figure 15

Growth Mindset Total Prediction Residuals Histogram

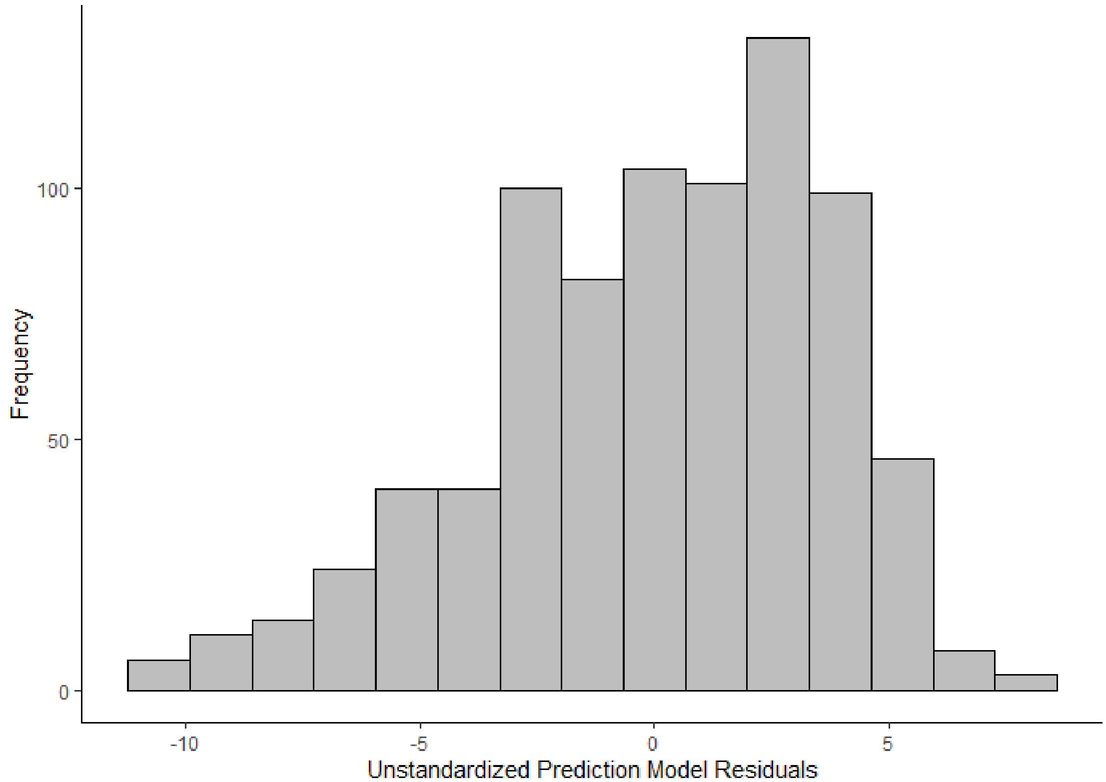
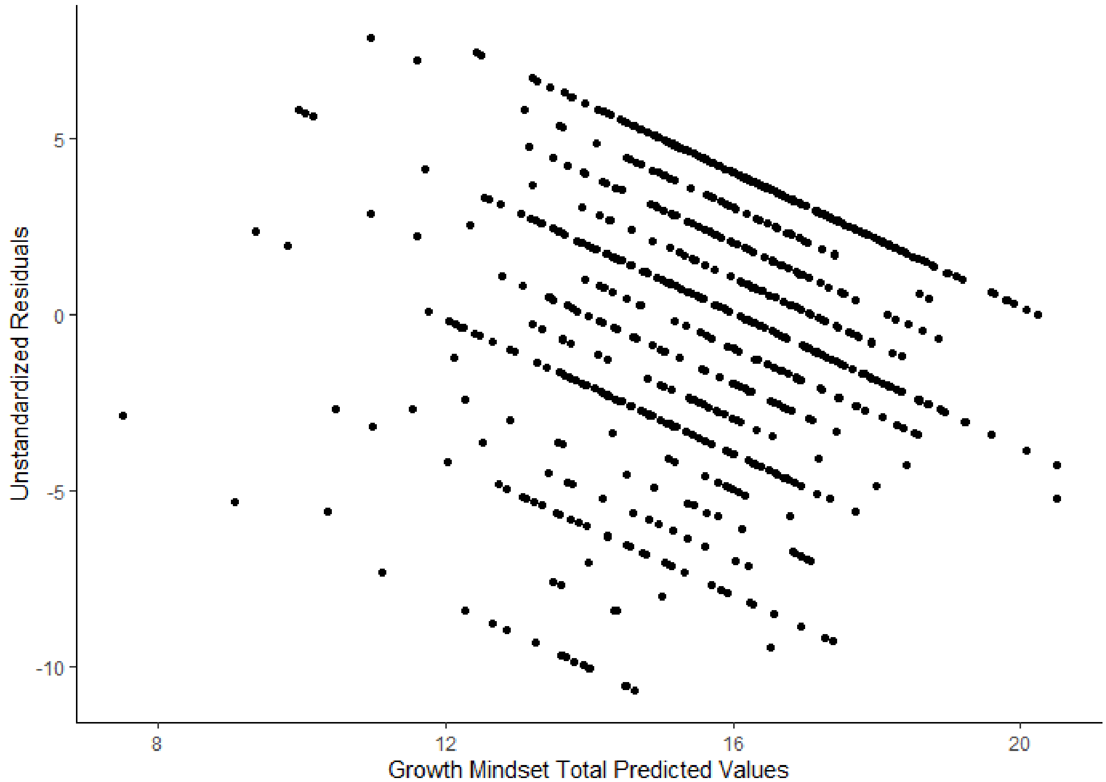


Figure 16

Growth Mindset Total Residual by Fitted Value Scatterplot



Appendix D

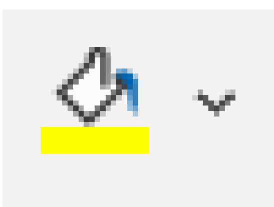
SSA Growth Mindset Rating Form Training Manual

This document is to guide SSA Agents in completing the SSA Growth Mindset Ratings Rubric. This rubric has been designed to help SSA Agents assess social emotional learning (SEL) competencies, a wide array of non-academic skills that individuals need in order to set goals, manage behavior, build relationships, and process and remember information, in the students they mentor.

Rubric Layout

The rubric is designed in three main sections: administrative information, SEL rating information, and rating context information.

Administrative Information



This section is highlighted with blue cells: Student Name and Agent Name. This is where all student mentees assigned to the SSA agent will be listed. If the name of a student you mentor is missing, please add it and highlight that row. To highlight the row, select the row number to the left of the sheet (e.g., 9) and go to the Fill Color button (displayed to the left). Choose a yellow color to highlight the row.

SEL Rating Information

This section is highlighted with green, orange, and yellow cells: rating scale, the SEL definition, the SEL competencies, and the SEL score. The green cells highlight the rating scale for the following SEL competencies (provided in the yellow cells). The ratings follow a five-point scale detailed in the table below:

Rating Value	Rating Label	Rating Example
1	Strongly Disagree	Through your interaction with the student, you perceive the student embodying the opposite of this competency.
2	Disagree	Through your interaction with the student, you perceive the student embodying the opposite of this competency, but you can think of a few instances where they did embody the competency
3	Neither Agree nor Disagree	Through your interaction with the student, you perceive numerous examples of them embodying this competency and the opposite of this competency; all of these examples blur together and you do not know which they exhibit more.
4	Agree	Through your interaction with the student, you perceive the student embodying this competency, but you can think of a few instances where they did not behave this way.
5	Strongly Agree	Through your interaction with the student, you perceive the student embodying this competency and cannot think of any instances where they did not behave this way.

The orange cell provides the definition of the SEL concept being assessed to provide context. For example, if Growth Mindset is being assessed this orange cell would be: "Growth Mindset: "In a growth mindset, people believe that their most basic abilities can be developed through dedication and hard work – brains and talent are just starting points." (Dweck 2015)."

The yellow cells break down this SEL concept definition into competencies. Each competency is to be rated according to the rating scale. The scores for each competency are then automatically combined in the SEL score cell to provide that student's SEL score for future analysis.

Rating Context Information

This section is highlighted with purple cells: provide rationale for the overall score, confidence level score, and confidence level explanation. This section details the rationale the SSA agent used to determine the scores they provided for the student. This rationale can help determine consistency between raters. Additionally, the SSA agent will rate their confidence in the accuracy of their score. The ratings follow a five-point scale detailed in the table below:

Rating Value	Rating Label	Rating Example
1	Very Unconfident	You do not feel you have had enough time to get to know the student to accurately rate them due to limited time partnered with them and/or lack of response/interaction from the student
2	Unconfident	You do not feel you have built up enough rapport throughout your time knowing the student that you relied on comparing this student to other students of yours in determining their ratings rather than using your previous conversations with this student
3	Neither Unconfident Nor Confident	You are unsure if you have built up enough rapport throughout your time knowing the student and made an estimated guess based on your previous conversations with this student and comparisons with other students of yours
4	Confident	You have built up enough rapport throughout your time knowing the student and made an informed estimate from examples and content of your previous conversations; no comparisons with other students were made.
5	Very Confident	You have built up enough rapport throughout your time knowing the student that your previous conversations easily portrayed examples to make ratings of each of these competencies or the opposite of these competencies; no comparisons with other students were made.

What To Consider When Rating

When making your ratings, you should consider all of your previous interactions with the specific student. Based on your text exchanges, have they provided any relevant examples where one of the listed competencies applies? When you have approached topics that relate to the SEL concept being assessed (e.g., Growth Mindset - persistence in face of challenges or improving performance with practice) does the student engage in the conversation or do they stop responding?

These ratings should be made independently; please do directly ask the student to rate themselves or compare ratings with other SSA Agents. Additionally, a high or low score is not necessarily better. Accurately rating each student can help align follow-up interventions and support structures to benefit the student's long-term development.

Furthermore, every student is not expected to be rated at the same time. You can wait to build up more rapport with a student before rating them. This rapport can be based on number of interactions or time spent mentoring them. This sense of rapport will be reflected in your confidence ratings.

Be Aware of Common Rating Errors

When rating your students, be aware of the following common rating errors:

Central Tendency Errors

These errors occur when raters only use the midpoint of the scale (i.e., 3).

Leniency Errors

These errors occur when raters only use the high end of the scale (i.e., 5). This error often occurs from raters wanting to look good, be liked, keep the peace, or maintain the culture. We want authentic ratings - these ratings do not relate to your performance as an agent.

Severity Errors

These errors occur when raters only use the low end of the scale (i.e., 1).

Halo/Horns Errors

Halo errors are when one good trait overshadows other traits, behaviors, actions, and beliefs. Horn effects are when one bad trait overshadows other traits, behaviors, actions, and beliefs.

Recency Errors

In this error, raters heavily weight their most recent observations.

First impression or primacy

This error is the opposite of recency errors; in this error, raters pay too much attention to initial experiences with the ratee

Similar-to-me

In this error, raters give more favorable ratings to ratees who are like themselves

Vita

Scott Blackburn has predominately been a career student, starting his academic career at University of North Carolina Wilmington in 2014. Over the next four years, he earned a bachelor's degree in psychology, bachelor's degree in criminology, minored in sociology, and graduated with university and departmental honors. Graduating with career uncertainty, he considered higher degrees in student affairs, management, and industrial-organizational psychology. Mixed results in graduate school acceptances and internal reflection led him to attend University of Florida from 2018-2019 to complete an accelerated master's degree in management. Transitioning to the full-time workforce having opportunities with Nabisco and PRG Real Estate, he obtained practical work experience, but still felt unfilled. Returning to his career uncertainty decision, he applied to graduate school again, getting accepted to attend Appalachian State University from 2020-2022, earning a master's in Industrial Organizational Psychology and Human Resource Management. Throughout the program, he worked on various professional projects, including leading educational analytics initiatives, a manufacturing safety audit, interning at Syneos Health resulting in a full-time offer, a performance management system evaluation of a regional manufacturing company, and assisting in the operations of a campus non-profit.