Training Wreck: A Comparative Analysis of Assessment Scores Before and After Mental Health Training

Emma Dantas[a], **Jose Luis Salinas Vargas**[a], **Katina Christensen**[a], **and Maya Magee**[a]

This manuscript was compiled on May 31, 2024

Medical professionals need to be well-versed in treating patients with intellectual and developmental disabilities (IDD). The purpose of this study is to evaluate the training modules used to equip medical students with the skills for best practice. Students took a pre-assessment that consisted of multiple choice, select-all-that-apply, and short answer questions about IDD and mental health which groups them into three cohorts: foundational, intermediate, and advanced. We conducted an analysis to compare medical students' pre- and post- scores across cohorts, evaluating their improvement. The advanced cohort saw improvement while the other two cohorts saw little to no change in scores after completion of the training module. We then examined whether there is correlation between prior knowledge and score improvement as well as whether there is correlation between students' sentiment about the training module and score improvement. The former has some correlation and is strongest amongst the advanced cohort, and the latter shows little to no correlation. Although there are some limitations, the training module does not appear to be very beneficial for the medical students learning to treat patients with IDD.

Medical students | START Training Modules | Mental Health and IDD |

Medical professionals' understanding of intellectual and developmental disabilities of their patients is vital for the success of patients' treatment. These are skills that can be taught to medical students through substantive training. Our project explores the improvement of medical students' understanding of intellectual and developmental disabilities using the National Center for START (Systemic, Therapeutic, Assessment, Resources, and Treatment) program, run by the Institute on Disability (IOD)/UCED at the University of New Hampshire (UNH). START is an evidence-based, cross-systems crisis prevention and intervention model for people with Intellectual and Developmental Disabilities (IDD) who have mental health needs. In order for prescribers who work with patients with IDD and mental health concerns to learn valuable techniques to aid "best practices," UNH has begun piloting training content with medical students. This content includes 6 hours of training based on the recently published Integrated Mental Health Treatment Guidelines for Prescribers in Intellectual and Developmental Disabilities (1). The training is designed to assess medical students' awareness, knowledge, and self-efficacy when treating patients with IDD (1). Participants also engaged with pre- and post- training surveys.

To begin, participants were asked to complete a "medical student knowledge" assessment before they began the training module, through which they were grouped into three cohorts. Next, participants completed the training modules, which were asynchronous and could be accessed at any time. All the content

Significance Statement

On average, individuals with intellectual and developmental disabilities (IDD) receive poorer quality healthcare than neurotypical individuals. This disparity can largely be attributed to a lack of training on the treatment of neurodiverse patients for student-physicians. The National Center for START Services created a training program which aims to close this treatment gap, and this paper analyzes the efficacy of this training in order to make recommendations for improvement and to steer the training in a positive direction with the goal of helping physicians provide the highest quality of care to all patients, regardless of their background.

Author affiliations: [a]Dartmouth College QSS20

in the training modules was organized in three levels: foundational, intermediate, and advanced. Each level included multiple modules and suggested readings. Upon completion, the participating medical students were asked to complete the same "medical student knowledge" assessment as well as complete course evaluations and surveys. These surveys had participants considering what they liked about the training and recommendations for how the training could be improved.

Previous work has shown the change in scores that occurs before and after training from a purely mathematical standpoint. Gitlin et. al's "Measuring Improvement in Medical Students' Understanding of Intellectual and Developmental Disabilities" highlights that across multiple choice questions and short answer questions, participants scores improved after training. They, however, did not make concrete distinctions between cohorts (2).

Our research aims to further assess the quality of the training modules. First, we consider how participants' scores change across the pre- and post- "medical student knowledge" assessment between each of the three cohorts individually. By looking at the foundational, intermediate, and advanced cohort score changes separately, we may be able to gauge how previous knowledge and different training module levels impacts participants' takeaways from the training. Next, we analyze how participants felt about the training through the use of natural language processing applied to their short answer post-survey responses. We conclude with an assessment of whether participants' feelings about the training modules correlate to their net score change on the pre- and post- "medical student knowledge" assessment. Moreover, we ask if medical students who had little to no score improvement have negative sentiment regarding the training and, vice-versa: if medical students who had more substantive changes to their scores have more positive sentiment regarding the training?

Data

To perform our analysis, we were provided data by the University of New Hampshire's National Center for START Services. As aforementioned, NCSS/UNH has begun piloting training content with medical students (residents, fellows, and second/third year medical students) as part of a strategic initiative to develop best practices for prescribers who work with patients with IDD and mental health concerns. The data we used is sourced from a survey the medical students took prior to engaging with such training (which we refer to as the pre-assessment or pre-training survey), a survey medical students took after completing the training (which we refer to as the post-assessment or post-training survey), and a module satisfaction evaluation survey where participants were able to give feedback after completion of the program. The earliest participant we have on record took the pre-assessment survey on 3/3/2022, and the last participant included in our data recorded their post-assessment answers on 2/3/2023. Thus, the time window we examine in our analysis of the training spans the period from March 2022 to February 2023.

The data came in the form of multiple .csv files that were generated from the results of both the pre- and post- "medical student knowledge" assessment and reflective survey. There is one file for the pre-assessment survey responses of all participants and three separate post-assessment files, one for each cohort. The questions on these assessments were multiple choice questions with one answer, multiple choice/select-all-that-apply questions, and short response questions. The pre- and post-assessment surveys asked medical students things like: "What are the most common mental health conditions experienced by people with intellectual and developmental disabilities?" and "How do you define positive medicine?" In addition to the questions concerning medical knowledge, the post-assessment also included a reflective survey evaluating the training modules which asked the participants questions like "What was the most helpful/what did you like about the training?" and "What recommendations do you have to improve the training?" For the multiple choice questions which only accepted a single answer, the string corresponding to the answer the respondent chose was placed in the number column corresponding to the question number. For the multiple choice questions which allowed the selection of multiple answers, the strings corresponding to each answer the respondent chose were turned into a single string with the individual answers separated by commas and placed in the column number corresponding to the question. Participants' responses to short answer questions were also recorded as strings in the questions' respective columns. Thus, all of the students' responses to survey questions were recorded as strings. After the fact, the assessments were scored using a defined system that, if necessary, allocated partial credit to the select-all-that-apply and short answer questions. Participants' scores were recorded in a separate column as a numeric digit.

The unit of analysis in the dataset and in our analysis is a single medical student — each row of the data set corresponds to a single medical students' responses to the questions in the survey. The data has various limitations. For starters, the survey allowed participants to drop out at any time. Because of this, many participants answered anywhere between 1-5 questions before quitting and leaving the remaining fields blank, resulting in an absence of responses for a variety of data fields. Additionally, only a small subset of participants completed everything in its entirety: the advanced cohort's post-assessment survey only has 29 complete responses; the intermediate cohort's only has 16; the foundational cohort's has 35. This relatively small sample size and the fact that it results from participants self-selecting whether to fully complete the assessments leaves lots of room for potential bias in our results. Additionally, the anonymity of respondents makes matching responses by participant identity very challenging, making it difficult to track the effects of the training on an individual basis and muddling conclusions about the training's efficacy as a result.

Methods

As mentioned above, the data was collected from the surveys and compiled into five separate datasets that we manipulated for our findings: the 'Med student pre assessment 2.8.23.csv' which contained every participant's survey and assessment results from before the training, the 'Foundational Post Assessment 2.8.23.csv', 'intermediate post assessment 2.8.23.csv', and 'Advanced post assessment 2.8.23.csv' which contained the results from after the training for each of the three cohorts, and finally the 'module satisfaction evaluations 2.8.23.csv' which contained responses regarding participants' feedback and feelings about the training.

Merging and cleaning. The original datasets were formatted in an inconvenient format, including unnecessary information in the first couple rows, so when loading in the data we used *skiprows* and a *lambda* function to omit the header information that might interfere with our data processing, making the data easier to read and work with. In order to begin our analysis and directly compare participants' scores, we needed to combine the pre- and post- assessments for each participant and further clean the data. We did this in our 01_merge_data_rename_columns.ipynb notebook by merging the datasets using a left join, matching respondents through the 'IPAddress' column that contained the IP Address of each participant as they answered the assessment. This process aligned the pre-assessment data with the corresponding cohort post-assessment data, allowing us to create three separate dataframes for each cohort. By using a left join, we merged in a way that only kept certain relevant columns and avoided keeping unnecessary rows with missing data that could negatively affect our analysis. Next, we renamed the relevant assessment score columns that had been affected by the merge—'SC0_y' became 'Pre' and 'SC0_x' became 'Post' for clarity—and then dropped any missing values from these columns to clean the data, forming three new dataframes: 'foundational_data', 'intermediate_data,' and 'advanced_data'.

Our next step was to transform the data to prepare it for analysis and to make visualizations. We converted the data for each cohort from wide format (individual columns for each assessment) to long format using the *melt* function. Switching to long format meant that we could more conveniently use *matplotlib* and *seaborn* functions to create plots that are segmented by cohort. In addition, we confirmed that all the score data in the 'Pre' and 'Post' columns was in integer format through explicit typecasting in order to ensure that it could be used in *matplotlib* plotting functions.

Next, we performed similar data preparations for the correlation plot and network visualization, assigning cohort labels to each participant and then combining the cohort-specific data frames into a single larger one in two different ways. The first way that we did this was by making use of a dummy labeling system for the network visualization phase. This process began by appending each cohort-specific dataset—advanced, intermediate, and foundational—with a new binary column. We then used the *pandas.concat* function to combine these datasets into one data frame. Next, we used '0's to fill in the binary indicators where there were missing values, ensuring that the data was complete and its integrity preserved. Next, we calculated the quartiles for the score column 'SC0' in the combined dataframe which we later used in our network analysis since the quartiles were used to map the distribution of connectivity with respect to score. Next, we created a cohort column directly from a condition-based assignment with *np.select*, that is where the cohort label (which was essentially the cohorts) was derived directly from the binary indicators that we created—'1' stands for the participant being in a specific cohort, '0' otherwise. This effectively labeled each participant to their corresponding cohort label. However, we later discovered an alternate, easier way of going about labeling the cohorts. Instead, we simply tagged each of the dataset labels with a 'Cohort' label and then merged the data. This not only simplified the steps of merging but also made it possible to easily grasp the cohort of a participant by viewing it in the merged dataset. By doing so, we could have merged the cohort labels directly without executing the extra steps of filling out our missing values and assigning the conditions post-merge.

Statistical Analysis. Our first element of analysis was a paired t-test in our 02_paired_ttest.ipynb notebook that explores the difference in scores from pre- to post-training within each of the cohorts. The paired t-tests were applied separately to the cohorts, and we chose to use this method because of its ability to compare repeated measurements on the same subjects, which makes it ideal for assessing the changes in scores within the same individuals.

Visualization. In a new notebook, 03_histogram_visualization.ipynb, we created three histograms, one for each cohort, that served as a visual representation of the distribution of participants' scores before and after the training. These histograms helped us assess a clear comparison between the spread and central tendency of scores before and after the training session within each cohort. We first transformed the dataset into a long format to facilitate the separation of 'Pre' and 'Post' training scores into two distinct categories within a single column. This manipulation was crucial as it allowed for the effective application of color coding to distinguish between the two conditions within the histograms. Making use of *seaborn* library's *histplot* function, we set specific parameters to change the appearance of the histograms to best fit our data and remain visually appealing. We chose to include a Kernel Density Estimation (KDE) line to display a smooth estimate of the score distribution to highlight the density and spread of the scores beyond the discrete bins.

These histograms were followed by a correlation plot, in 04_correlation_test.ipynb, that shows how scores correlate across the different cohorts. These comparisons provided a straightforward depiction

Table 1. Overview of the Analysis Sample

CSV	Number of Respondents	Date Range
1. Pre-assessment survey	100	Mar 2022 - Jan 2023
2. Foundational post-assessment survey	35	Mar 2022 - Jan 2023
3. Intermediate post-assessment survey	16	Mar 2022 - Oct 2022
4. Advanced post-assessment survey	29	Mar 2022 - Feb 2023

of potential trends or patterns in the data, helping us understand the overall impact of the training and how it differed for each cohort. To achieve this, we utilized seaborn's *regplot* function to fit and visualize a linear regression model between two variables which were the pre- and post- training scores. Within the 'data' parameter of *regplot*, we used a for loop to selectively include scores from each cohort. This method ensured that all three correlations could be in a singular plot, facilitating an easier comparison and interpretation of how the training impacted each group distinctly.

In our 05_network_analysis.ipynb notebook, we created a comprehensive network analysis to display the relational dynamics within the cohorts based on their assessment scores. The network visualization was constructed to map the connections between participants within the same quartile of scores, facilitating an understanding of the distribution and clustering of scores across different cohorts. We utilized the *NetworkX* library to create a graph where each node represented a participant, labeled by their score quartile and cohort. Nodes were connected if participants fell within the same score quartile, reflecting the potential interactions or similarities in score dynamics within the cohort. To enhance the visualization, we assigned colors to nodes based on their cohort, using a predefined color map. This not only made the network visually engaging but also intuitive to analyze, as the color-coded nodes provided immediate visual cues about cohort distribution. The visualization was then executed using the *Netwulf* library, which allowed for an interactive exploration of the network.

Our final element of analysis, in 06_sentiment_analysis.ipynb, was a sentiment analysis of how participants felt about their training and assessments. We imported the *SentimentIntensityAnalyzer* from Vader and loaded in the data from the 'module satisfaction evaluations 2.8.23.csv,' creating a pandas data frame 'satisfaction_df.' We merged the data frames that we had used in previous notebooks using a left join on the 'IPAddress' column and assigned cohorts to these new data frames before using *pd.concat* to combine them into a singular, larger data frame. Then, using Vader, we calculated the polarity scores of the 'What was most helpful?' column and assigned this to the 'sentiment_score' variable. Next, we calculated each participants' score improvement by subtracting their 'Pre' score from their 'Post' score and assigned this to the 'SC0 Improvement' variable. To plot the correlation between the two new variables we used matplotlib's *regplot* to make a regression of the correlation between positive sentiment and score improvement. Similarly, we created another regression plot that assigned the corresponding colors to the cohorts and separated the correlations of each of the cohorts in a singular plot so that we could directly compare whether the correlation between sentiment and score improvement was different in each of the cohorts.

Results

Based on a high-level analysis of the students' performance by cohort, it appears that the training resulted in little improvement between participants' pre-training and post-training scores. Figures 1 and 2 provide a direct comparison between how each cohort scored on the assessments before and after taking the training. For both the foundational and intermediate cohorts, overall performance seems to have declined as a result of the training, as post-training scores were consistently lower than their pre-training counterparts. Thus, figures 1 and 2 imply that the training was ineffective in educating participants who came into the training with little to no background knowledge on treatment of individuals with IDD.

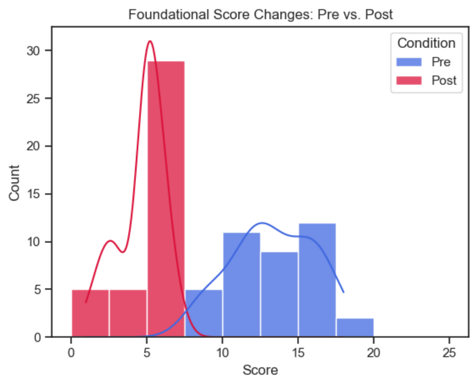

Fig. 1. Comparison of respondents in the foundational cohort's assessments before and after taking the training. We see a worsening of scores.

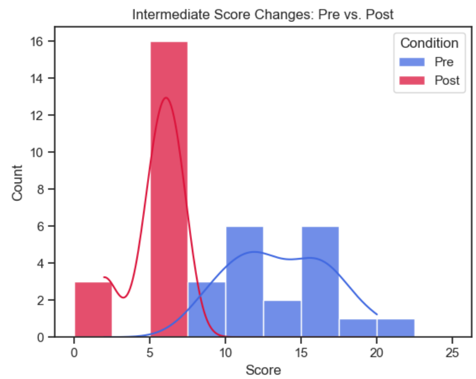

Fig. 2. Comparison of respondents in the intermediate cohort's assessments before and after taking the training. We see a worsening of scores.

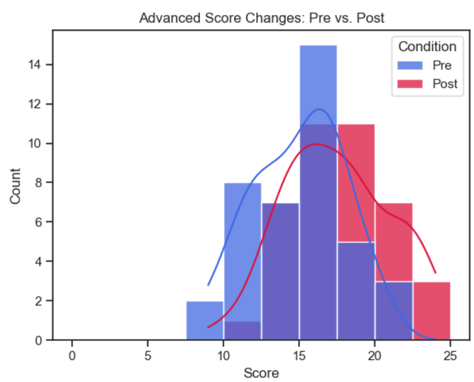

Fig. 3. Comparison of respondents in the advanced cohort's assessments before and after taking the training. We see an improvement in scores.

The initial analysis of the advanced cohort's performance is more promising—figure 3 illustrates that the advanced cohort outperformed their pre-training scores on the post-training assessment. In this case, both the minimum and maximum scores on the post-training assessment are higher than those from the pre-training assessment, and the frequency of high scores (scores within the range of 17-25) increased. We can clearly see that the distribution of scores on the post-training survey has shifted rightward (in the positive direction) as compared to the distribution of scores in the pre-training survey, further emphasizing that the advanced cohort's understanding improved as a result of the training.

In order to further explore relationships between participants' pre- and post-training scores, we conducted a correlation test, visualizing the correlation between pre-training and post-training scores across different cohorts using regression plots in figure 4. In this case, the advanced cohort is the only cohort in which there is demonstrated correlation between pre-training and post-training scores—the green line shows positive correlation with a noticeable upward trend, indicating that higher pre-training scores are associated with higher post-training scores within this group. The red and yellow lines which represent the foundational and intermediate cohorts, respectively, show less pronounced trends, suggesting weaker correlations between pre- and post-training scores. For the bottom two cohorts, it appears that how participants perform on the pre-training assessment does not have a large impact on their performance on the post-training assessment. Thus, our initial analysis strongly suggests that the training's efficacy as an educational tool is rather limited and that cohort placement plays a key role in whether the training is remotely effective.

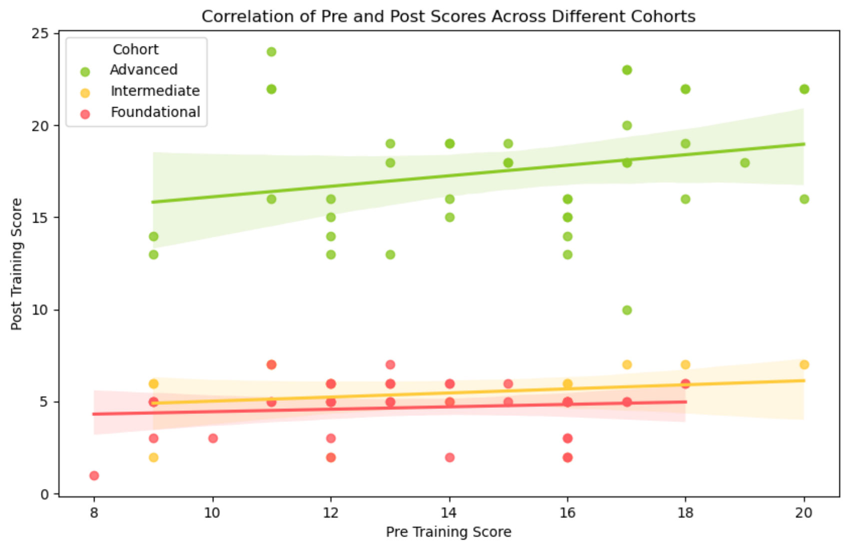

Fig. 4. Tracking the correlation between pre-training and post-training scores across the three cohorts. Only the advanced cohort demonstrates clear positive correlation.

To confirm our hypothesis that the training's effectiveness heavily depends on an individual's baseline knowledge and cohort placement, we broke up the participants and their scores into performance quartiles and used network visualization techniques to investigate the distribution of scores across different cohorts and quartiles. Our results were very much in line with our hypothesis—figure 5 highlights that, as expected, the quartile in which participants scored largely correlates to their cohort although there were a few outliers who significantly outperformed or underperformed compared to the rest of their cohort. For instance, the top quartile of performers consists entirely of participants from the advanced cohort. The second-best quartile was comprised largely of participants from the advanced and intermediate cohorts; however, there were a couple participants from the foundational cohort who scored in this range, providing us with a glimmer of hope for the training's effectiveness for individuals with little background knowledge. The third-best quartile was mostly made up individuals from the foundational cohort with a few participants from the intermediate cohort scoring in that range, and the bottom quartile was mostly composed of individuals from the foundational cohort with some intermediate participants and a couple advanced participants also scoring in that range.

Additionally, the network visualization illustrates that, based on the varying sizes of the clusters, the distribution of performances is rather extreme. The majority of participants fall into either the top or bottom quartile, making those clusters larger and more dense than the clusters representing the middle two. Additionally, the smaller, middle quartiles are also characterized by looser connections between participant nodes, suggesting more diversity in scores which results in more dispersed clusters. Therefore, the overarching trend is that most participants tend to do very well or very poorly with few falling in the middle range, implying that there is a polarity of understanding and comprehension amongst individuals who participate in the training.

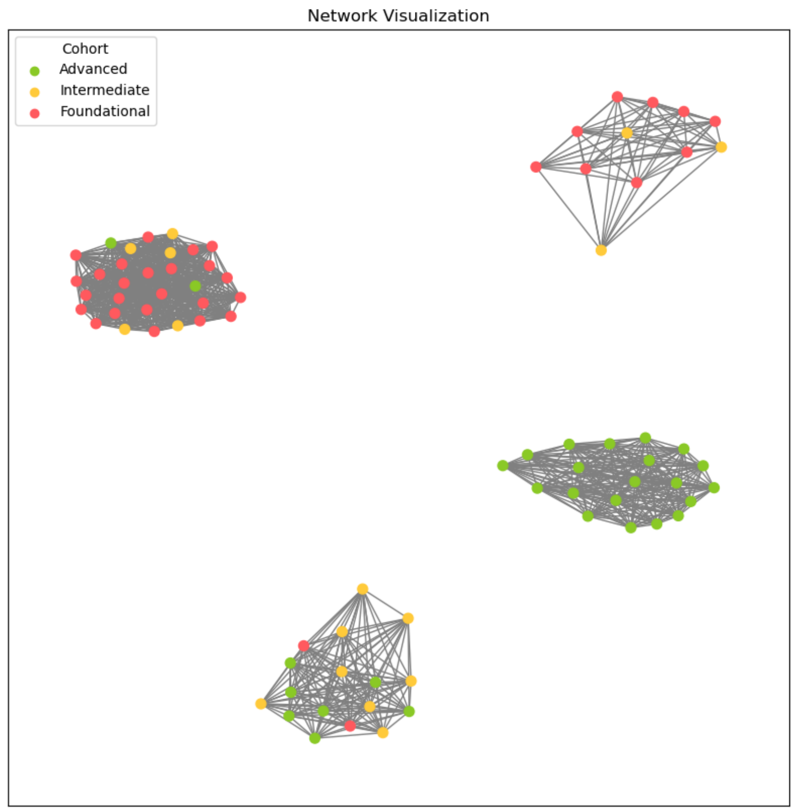

Fig. 5. The above network visualization explores the compositions of and connections among score quartiles. For clarity's sake, the top 25 percent is the rightmost cluster; the cluster in the 25-50th percentile of performers is on the bottom; the group in the 50-75th percentile of performers is the uppermost cluster; and the leftmost cluster represents the bottom 25 percent of performers.

Next, we analyzed how participants felt about the training with the goal of assessing whether participants' feelings about the training correlated to their net score change on the pre- and post-training assessments. In figure 6, we visualized the results of NLP sentiment analysis on a short answer question which asked participants "What was most helpful?" This analysis helps us gauge whether score improvements relates= to positive sentiment about the training, highlighting potential biases in participants' reviews. Our first step was to look at these results across aggregated cohorts. Even though there appears to be a slightly positive correlation between the two variables, the regression returns a line of best fit for which the score improvement is negative for all values. Thus, when cohorts are combined, it is difficult to make a clear case about whether there are significant connections between how participants feel about the training and their performance.

We then looked at the sentiment analysis when participants were broken up into their respective cohorts in figure 7, as we believe that looking at results cohort by cohort is more helpful in evaluating the efficacy of the training. We found that for the advanced cohort, students typically show positive score improvement regardless of how positive their sentiment was. However, students in this cohort with positive sentiment show a larger increase in scores, as the advanced cohort seems to broadly follow the behavioral pattern we expected participants to demonstrate—that participants who

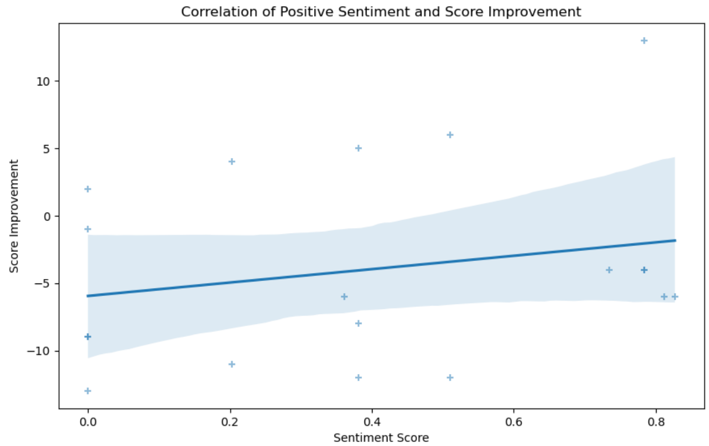

Fig. 6. Comparing the impact of course satisfaction on score improvements across cohorts. The x-axis tracks participants' positive sentiment in their response to the question "What was most helpful?" Whether there is a significant connection between how participants feel about the training and their performance is not entirely clear from this combined graph.

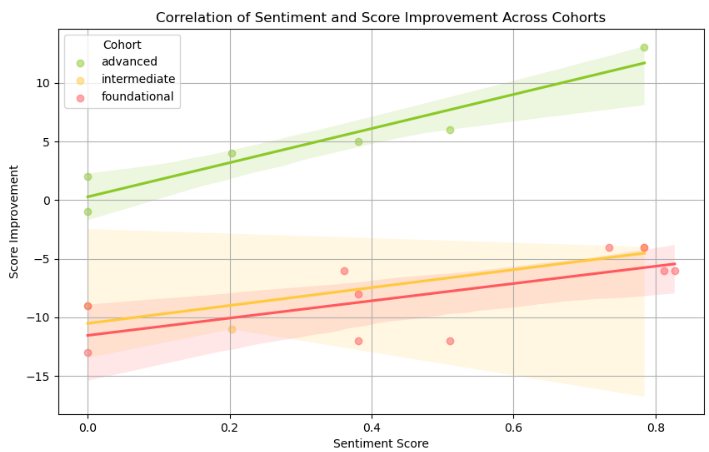

Fig. 7. Showing the impact of course satisfaction on score improvements by cohort. The advanced cohort is the one which most closely seems to follow the behavioral pattern we expected participants to demonstrate—that participants who improved the most as a result of the training would be the most positive in their reviews.

improved the most as a result of the training are the most positive in their reviews. In the advanced cohort's case, there seems to be a significant positive correlation between sentiment and score improvement.

However, the advanced cohort is the only one which demonstrates this trend in a significant way. Both the intermediate and foundational cohorts exhibit score decline, regardless of sentiment. Even though those with positive sentiment might perform slightly better, our takeaway is that there is little relation between positive sentiment and score improvement for these bottom cohorts since regardless of sentiment, participants in these cohorts tended to experience score decline. These results reinforce our previous analysis which holds that the training, in its current state, is extremely limited in its power to successfully educate participants in the foundational and intermediate cohorts. They also suggest that how participants feel about the course is not as closely intertwined with its educational success as one might expect before exploring the data.

Discussion

In conclusion, even though participants generally felt positive about the training modules and gave satisfying reviews, we found that the training modules had limited efficacy in truly improving medical professionals' understanding of IDD and how to better treat patients with IDD. We found a crucial asymmetry in terms of pre- and post- assessments: cohort matters significantly in both scores and sentiment. The training appears to be most effective for students who came into the training with some level of background knowledge concerning treating patients with IDD as the advanced cohort was the only one which saw marked improvement in scores after their training—participants classed in the foundational and intermediate cohorts saw little to no improvement in scores. Thus, the National Center for START Service should seek to improve the training, especially targeting improvements in its power to educate medical professionals for whom the training is their first introduction to interaction with patients with IDD.

One of the biggest limitations within our study is the sample size. For starters, the background information on the survey stated that you could drop out of the pre- or post-survey at any time. Because of this, it appears within the data that while some participants answered all the questions, others answered anywhere between 1-5 questions before quitting and leaving the remaining fields blank. While the data set may seem extensive, only a small subset completed everything in its entirety which was made obvious to us when we cleaned the data. The post-training results for the advanced cohort only has 29 responses; the intermediate cohort has 16; the foundational cohort has 35. With this small sample size it is difficult to draw concrete correlations.

Additionally, while working with the data, we discovered the difficulty of matching participants' between pre- and post- survey. We did not have access to participants' names because the Center for Social Impact at Dartmouth College, who we obtained the data from, has to protect the anonymity of the participants. Because of this, we initially attempted to link participants between the response ID column. However, it turned out that participants were given two unique response IDs: one for the pre-survey and one for the post-survey. We then pivoted to link participants across pre- and post-survey by IP address. The issue with this was twofold. First, a participant may have taken their pre-assessment in one location and their post-assessment in another. This would mean that their pre- and post- IP addresses were unique and therefore unmatchable. As a result, for the consistency of our analysis, we lost both of these data observations, further reducing our sample size. Secondly, there were duplicate IP addresses for unique participants within the pre- and post-assessment data. This meant that, even though we attempted to connect the IP addresses in a one-to-one fashion, we cannot guarantee that this occurred. Links between pre- and post- assessment IP addresses may have been made across participants randomly: person A's pre-assessment results could be linked to person B's post-assessment results because of a duplicate IP address. Once again, we may lose a pre- or post- assessment observation in our already small sample size due to this issue, making it difficult to pinpoint concrete correlation when drawing conclusions. Although we acknowledge these limitations amongst others, our study finds that the training module does not appear to be very beneficial for the general medical students learning to treat patients with IDD, and we urge the National Center for START Service to improve the training.

ACKNOWLEDGMENTS. We would like to thank the Dartmouth Center for Social Impact for connecting us with this study, and we would like to thank Professor Chang for providing us with the tools to work on this project and for his help along the way.

1. Center for START Services, Intellectual and developmental disabilities and mental health prescribing guidelines (2024) Accessed: 2024-05-28.
2. S Gitlin, K Hamann, R Williams, Measuring improvement in medical students' understanding of intellectual and developmental disabilities. pp. 1–11 (2022).