Measuring Improvement in Medical Students' Understanding of Intellectual and Developmental Disabilities

Saige Gitlin[a], Kayla Hamann[a], and Rachael Williams[a]

[a] Dartmouth College, Hanover, NH, 03755

This manuscript was compiled on November 23, 2022

This paper examines the outcome of the The National Center for START Services' training modules for medical students caring for patients with intellectual and developmental disabilities (IDD). People with IDD represent a not-insignificant portion of the population. By the nature of many IDDs, these patients are particularly vulnerable to having their needs mistakenly overlooked by providers if not trained properly. Through data analytics and natural language processing (NLP) techniques, the pre- and post-assessments of the medical training modules are quantified and graded. Across all short answer and multiple-choice questions, an increase in performance is seen which represents the effectiveness of the training modules. The relative increase can be examined to determine where the training modules can be further optimized to improve medical students' training outcomes even more.

Intellectual Disability | Developmental Disability | Medical

Currently, most physician training is specifically tailored for providing adequate healthcare for and treatment of neurotypical individuals. However, accessibility advocates are beginning to draw attention to the importance of training physicians and medical professionals on how to properly community with and treat patients with intellectual and developmental disabilities (IDD) due to the large number of individuals that have these conditions. For example, according to the Center for Disease Control (CDC), nearly 17 percent of children between ages 3 and 17 have one or more developmental disabilities. Developmental disabilities are categorized as "impairments in physical, learning, language, or behavior areas." (1) To be classified as developmental disabilities, these conditions must arise before individuals reach the age of 22 and typically last for a lifetime. The most common developmental disabilities are autism spectrum disorder and cerebral palsy. On the other hand, intellectual disabilities are a specific type of developmental disabilities that arise before the age of 18 and are defined as, "significant limitations in both intellectual functioning and in adaptive behavior, which covers many everyday social and practical skills." (1) Individuals with intellectual disabilities typically have an IQ score of 70 or lower while the average American IQ is 98 and 68% of Americans have an IQ between 85 and 115(1). According to the American Academy of Pediatrics, intellectual disabilities affect approximately 1-3% of the population(2).

Significance Statement

On average, individuals with intellectual and developmental disabilities (IDD) receive poorer quality healthcare than neurotypical individuals. In many ways, this disparity is attributed to a lack of training on the treatment of neurotypical patients for student-physicians. Developing effective training on the treatment of patients with IDD is an essential step towards eradicating this disparity. This paper provides an analysis on the quality of the training program created by The National Center for START Services on treating patients with IDD. Through analyzing the outcomes of this training, we can make recommendations for improvement and steer the training in a positive direction. Ultimately, the development of a successful training program can help physicians provide the highest quality of care to all patients.

Thus, even though individuals with intellectual and developmental disabilities (IDD) make up a sizable component of the population, when it comes to the healthcare of individuals with IDD there is still systemic inequity. In particular, the CDC reports that people with IDD have poorer health outcomes, shorter lifespans, are often victims of various forms of assault, and are more likely to have unmet health needs than neurotypical individuals.(1)The CDC also notes that individuals with IDD make up a significant proportion of potentially avoidable emergency room visits each year, suggesting that people with IDD often do not have their health needs addressed until they are severe. This delayed care is not only dangerous for individuals with IDD, but also puts unnecessary financial strain on the hospital system and the patient. (1)

In an effort to alleviate some of these disparities, The National Center for START Services at the University of New Hampshire, which aims to provide support to people with IDD, piloted a training program built from their recently published 'Integrated Mental Health Treatment Guidelines for Prescribers in Intellectual and Developmental Disabilities' and aims to teach medical students best practices for caring for and interacting with patients with IDD. This asynchronous training program contains up to six hours of information on treating patients with IDD and includes lessons on neurodiversity, culturally competent care, and positive medicine. At more advanced levels, the training addresses more complex issues such as how to diagnose and treat mental health disorders in IDD patients. Prior to the training and after each two hour chunk of training, participants were asked to answer multiple choice and short answer questions about treating and engaging with patients with IDD. This paper aims to analyze the pre and post training survey data to identify strengths of the training as well as to assess if the START Services training was ultimately effective in improving medical students' understandings of IDD patients and their treatment. More specifically, this paper aims to determine which topic areas students improved the most in as well as which they improved the least in after training to identify areas where the training can be improved and areas where the training is achieving high efficacy and retention.

Data. To perform this analysis, we were provided data by the University of New Hampshire's National Center for START services. The data is sourced from a survey medical students took prior to engaging with the aforementioned training material (will be referred to as the pre-training survey) and a survey medical students took after completing the full six hours of training (this is referred to as the advanced post assessment by the data provider but will be referred to as the post-training survey in this paper for clarity going forward). The surveys were taken between March 2022 and October 2022.

Basic Characteristics of the Data Set

DataSet	Number of Eligible Observations	Dates Covered
Pre-Training Survey	49	March 23 - October 10, 2022
Advanced Post-Training Survey	16	March 30 - October 8, 2022

Table 1. This table provides information on the number of observations in each dataset as well as the time frame they cover.

In both surveys, students answered three multiple choice questions that allowed respondents to select only one answer, seven multiple choice questions that allowed respondents to select multiple answers, and three open-ended short answer questions where respondents provided a personalized typed response. The data was provided in multiple .csv files, one for the pre-training survey and one for the post-training survey, with columns corresponding to each of the questions in the survey.

For the multiple choice questions which only accepted a single answer, the string corresponding to the answer the respondent chose was placed in the column corresponding to the question. For the multiple choice questions that accepted multiple answer selections, 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 respective column. Lastly, for the short answer questions, the hand typed responses given by the students were provided as strings in the respective column. Thus, all of the responses were provided as strings. The questions asked the medical students things like: "What are common mental health conditions of people with IDDs?", "Why is the recognition of neurodiversity important?", and "What are best practices in communicating with patients with IDD-MH and their families?" in order to identify whether the respondents are grasping the best practices the training is attempting to convey. As an aside, we were provided the solutions to all of the questions in a separate word document which we converted into a .csv file for future use.

The unit of analysis in the dataset and in our analysis is a single medical student as each row in the data set corresponds to one medical students' responses to the questions in the survey. In the pre-training survey data set derived from the survey students took before viewing the training material, there were 62 total respondents. However, only 49 of those respondents actually consented to participate in the study so our pre-training survey dataset consists of 49 observations. In the post-training survey, the survey students took after completing all six-hours of the training, there were only 16 respondents. We chose to use the last post-training survey (the advanced post assessment) as our post-training metric because it was the last survey the medical students took, so it would show the effects of the entire training as a whole. Additionally,the advanced post assessment was the only survey taken throughout the training that included all of the exact same questions as the pre-training survey and thus it is the only one we could equally compare to the pre-training survey. However, our choice to use the advanced post assessment as our only post-training metric was definitely a limiting one. As mentioned before, the pre-training survey had 49 eligible respondents and this number dropped to 16 respondents as the training progressed as the medical students were not required to complete the full six-hour training. This is such a small number of respondents in our post-training survey (16) that it will undoubtedly bias our metrics of improvement to represent the experiences of these few students and will leave our metrics open to being biased by outliers.

Methods. The first step in our analysis process was to clean the data and we focused our initial cleaning efforts on the multiple choice questions. The first step in our cleaning process was to simply subset the pre-training survey dataset to include only those respondents who consented to participate in the study, as mentioned above (code 02,line 110-112). The second step in our process was to rename the column names in the post-training survey dataset so that they matched the column names in the pre-training survey dataset (code 02, line 196-207). Next, we used regular expressions and the 're' package in python to clean up some of the strings in the respondents answers (code 02, line 25-30). We created a regex pattern for each question whose answers included a comma within the answer and replaced that part of the answer with the same string but with the comma removed. For example, 'talking to themselves, monologue' became 'talking to themselves monologue'. We did this for questions 31, 36, 34, 39, and 12 as these were the only questions that accepted multiple answer selections and had commas within the answers. Additionally, we used regular expressions to fix some differences between the way answers were written on the pre-training survey and the post-training survey. In particular, we identified that on question 39, the wording and capitalization of two of the answer choices were different between the two surveys so we corrected these for consistency.

Not only did we do basic string cleaning of the multiple choice answers, but we also decided to create and utilize a mapping system (02, line 164). When we converted the solutions from the provided word document to a machine-readable .csv file, each answer choice was given a row in the dataset. We then assigned each answer choice for each question a number so that the numbering scheme was unique to each question but not to the document as a whole and placed these numbers in a new column. For example, if a question had four answer choices, they would be mapped to the numbers one through four in the order they were presented on the surveys. After we created the mapping, it was time to apply it to the pre-training survey data set and the post-training survey data. First, for the columns which corresponded to a question that accepted multiple answer selections, we looped through the rows of each data set and split the answers with a comma as the delimiter; this was because the questions with multiple answers had the answers respondents chose presented as one string in the cell with the individual answers choices concatenated with commas in between them. Splitting the strings produced a list of strings in each row of each column, which we then looped through and used a map function to apply our mapping dictionary. Thus, in each cell there was now a list of integers which corresponded to the answer choices a respondent chose for a specific question.

Once the multiple choice responses were cleaned, it was time to measure the students performance on the questions before and after taking the training in order to identify whether or not the training was effective for the topics covered in the multiple choice questions. For the questions which only allowed respondents to select one answer, it was easy to measure performance. Students were given 1 point if they chose the correct answer and 0 points otherwise. We then calculated how many students got the question right and divided this by the number of students to get the overall group score for each question on both the pre-training survey and the post-training survey (02, line 360-374). The results of this analysis are displayed in Figure 1. Additionally, we measured the difference between these scores on the pre and post training survey and this quantification is displayed in Table 2.

Measuring the students' performance on questions which allowed students to select multiple answers was trickier as we wanted to penalize students for selecting incorrect answers as well as give them partial credit for at least selecting some correct answers. To achieve this, we created a custom scoring system that would grant students 1pt if they only selected all of the correct answers, a positive score between 0 and 1 if they selected more correct answers than incorrect answers, a score of zero if they selected an equal number of correct and incorrect answers, and a negative score if they selected more incorrect than correct answers. To calculate this score, as seen in equation 1, we first checked how many of the items in the list of answers a student selected were in the list of correct answer choices then subtracted the number of items in their response list that were not in the correct response list and divided this value by the total number of answer choices. Next, in equation 2, we calculated the maximum score a respondent could get on each question by taking the number of correct answer choices and dividing this by the total number of answer choices. Lastly, in equation 3, we took the respondent's score and divided it by the maximum score to normalize their score. Once we calculated the score for every respondent for every question, we then calculated the mean score on each question to identify how the group did as a whole on the pre-training survey and the post-training survey. These raw scores are plotted on figure 2 and we also quantified the difference between these averages in Table 2(code 04, 106-126).

$$respondentscore = \frac{\#correct - \#wrong}{\#options} \quad [1]$$

$$maxscore = \frac{actual\#correct}{\#options} \quad [2]$$

$$overallscore = \frac{respondentscore}{maxscore} \quad [3]$$

The short answer questions required additional cleaning strategies to prepare them for an analysis via natural-language processing (NLP). The file 03_text_cleaning.py is our module to make every piece of cleaning its own function, called by one clean_text function (Lines 31-42). The sample short answer responses were made lowercase (Line 32), cleaned of punctuation (Lines 54-57), excess spaces (Lines 34-45), and stop words from the nltk toolkit (Lines 60-67), and lemmatized (Lines 73-105). Lemmatization is similar to stemming in that both stemming and lemmatization consolidates many words down to a a root word, however lemmatization retains more context-specific information by taking in the part-of-speech (POS) tag. A classic example where stemming misses context is turning "caring" to "car" instead of the proper root word "care" like lemmatization does, because it knows that "caring" is a verb POS. Given that the short answer responses use just a few words to embed the meaning behind a central idea, lemmatization seemed like the appropriate choice for this use case.

The student's short answer responses underwent the same preprocessing as the sample short answer responses, but with an additional functionality to clean typos as well. Like all spellcheckers, the module used for correcting spelling errors, TextBlob, does not guarantee a correct output however text fields are cleaner on average after being corrected for errors than before. Once these steps of preprocessing have been performed, the text was ready to create a bag of words for grading.

The sample short answers responses were used to create a bag of words to grade the student short answer responses against. The lemmatized, cleaned words from the student answers are compared to all the words in the bag for a given question. The proportion of words which appear both in the student short answer responses and the sample answer response determine whether or not the student's response is graded at full credit (1 point), partial credit (0.5 points), or no credit (0 points). However, given that the course materials only provide one sample short answer response, this method could penalize students

Hence, the existence of synonyms in natural language must be taken into account. Many dictionaries exist online which have synonyms listed, so this resource could be leveraged to make the short answer grades more robust. Since thesaurus.com specializes in synonyms and antonyms of words with simple HTML to display them, this website made the best choice to scrape for synonyms to the words in the sample answer key. BeautifulSoup makes fetching these words from the HTML easy since the desired synonyms can be found at class="css-1wgnrap e1ccqdb60".

We were interested in evaluating how our natural language processing model was performing, so we hand graded the responses as well and will compare the results later in the 'Results' section. Given there were less than 200 responses to grade by hand, this was a feasible way to check the accuracy of our model. In order to hand grade the short answer responses, we used the rubric provided by The National Center for START Services. All short answer questions were graded for consenting participants in the pre-assessment and the advanced post assessment. On the rubric, participants could receive 0, 0.5 or 1 point for their answer. Full credit was awarded when students met all parts of the answer, while half credit was awarded if they only mentioned parts but not all of the correct answers. In the hand grading process, if it was unclear whether or not an answer was worthy of credit, the score was always rounded down for the purpose of consistency. It is also important to note that answers were not graded solely according to the keywords provided in the answer key. For example, the answer to Question 40, "How do you define positive Medicine," was "Positive medicine is the focus on both physical and mental wellbeing that promotes a sense of hope, and allows the person to thrive." A respondent who answered the question with the following answer, "I would define positive medicine as reframing discussions about health with a patient by focusing on strengths as well as identifying aspects of a patient's life that they personally find distressing, rather than only focusing on caregiver's concerns," was awarded full credit. Although they did not directly state the word 'wellbeing,' the student spoke about focusing on the patient's concerns to promote better treatment. While they did not use the word 'hope,' they spoke about focusing on a patient's strengths, which indicates they believe positive medicine is about a more hopeful outlook. Through using this approach, the hand-grading system focused on evaluating the deeper meaning of the answers and was able to evaluate responses beyond the simple inclusion of buzzwords.

After grading, the average score was calculated for each question in the pre assessment and then for each question again in the advanced post assessment using excel. Averages were used to account for the different number of participants in the pre and post-assessments and to allow for a more accurate comparison of the two surveys. The data from these averages were then imputed into a dataframe and a plot was created in python(code 04, line 180-219).

Results. At a high-level analysis of the students' performance, it appears that the students improved on the topics covered in the multiple choice questions as they had greater accuracy on all 10 multiple choice questions after taking all 6 hours of the training. If we look at Figure 1 and Figure 2 we can see a direct comparison of the average score of the group on the multiple choice questions before and after taking the training. For every question, the post-training survey bar is higher, or more positive, than the pre-training survey.

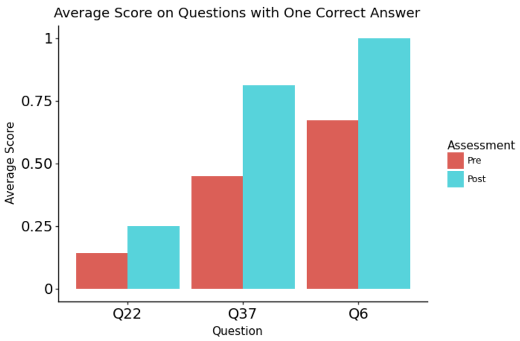

Fig. 1. Respondents performance on the survey questions that accepted one answer before and after taking the training. We see improved accuracy on all of the questions.

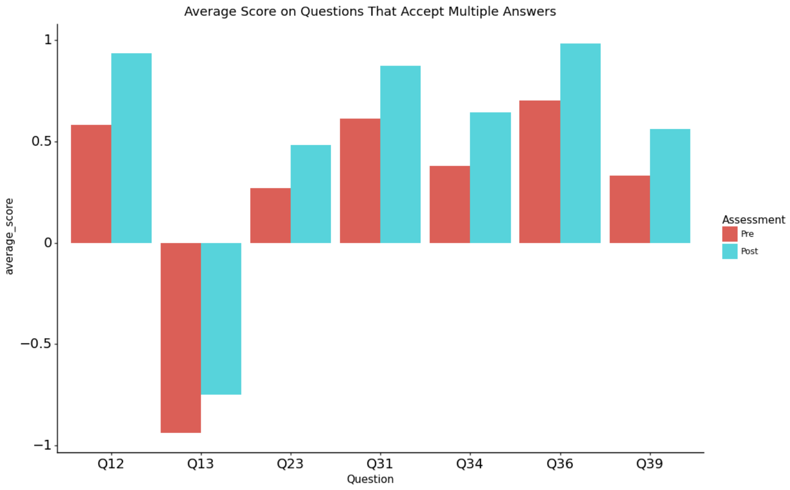

Fig. 2. Respondents performance on all of the survey questions that allowed respondents to select multiple answer choices. The height of the bars corresponds to the average score of the group as a whole and we notice the post-training survey averages are all higher than the pre-training survey averages.

We provide a more nuanced plot for a single question in Figure 3 where we display the entire distribution of scores for a sample question with multiple correct answers. In this granular view, we can see how 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 which signifies that the students understanding improved on this question. While Figure 3 is helpful for identifying how many respondents are getting each score as well as understanding the students performances more closely to identify stragglers or high achievers, it is hard to quantify improvement this way which is why we utilize the average score on Figure 2. This average score is the same score we use in Table 2 and we caution that while it is a helpful metric for efficiently and concisely identifying whether students are improving after taking the training, the average is highly susceptible to being biased by outliers which is a concern particularly in the post-training survey given it has only 16 observations.

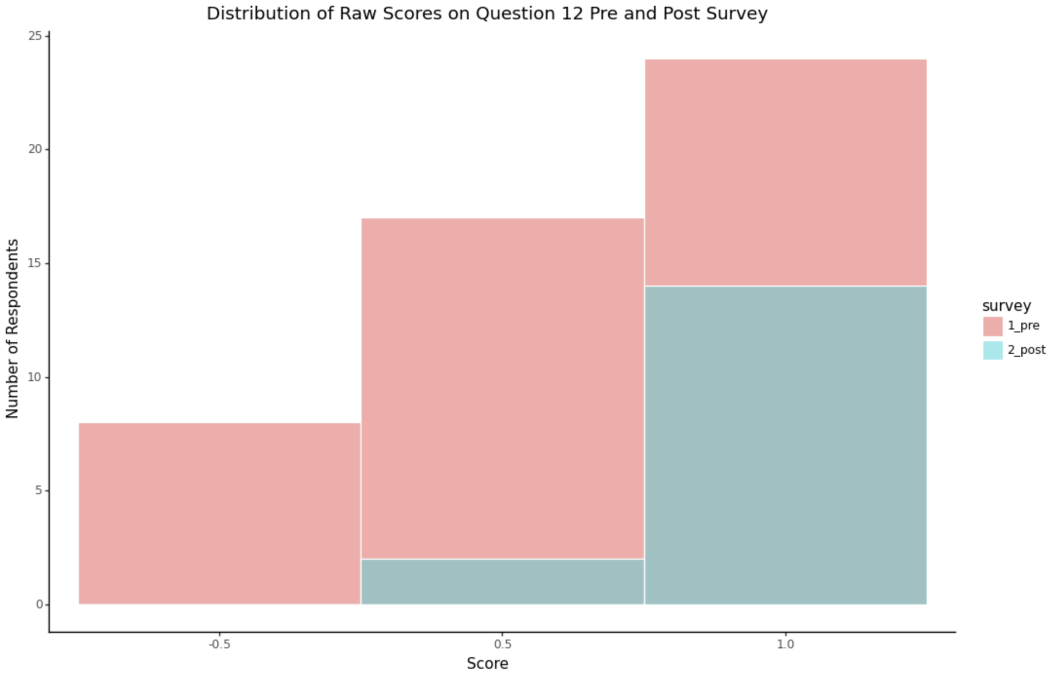

Fig. 3. The number of respondents that got each score on question 12. Question 12 asks: Which of the following are best practices in communicating with patients with IDD-MH and their families? Select all correct answers. It had four answer choices and two of them were correct. We see a distinct rightward shift in the post-training survey scores compared to the pre-training survey scores.

Next, we look at Table 2 to provide us a more simplified and understandable look at the results. Again, we can identify that all of the multiple choice questions (those with type 'One MC' and 'Multi MC') have a positive change meaning that the medical students' understanding of IDD improves after having taken the training. Specifically, we draw attention to question 6, question 37, and question 12 which all experience a jump in accuracy greater than 30% between the pre and post training surveys. Question 12 in particular is interesting because it had multiple correct answers which provides the opportunity for respondents to select extraneous answers versus being restricted to one answer choice. These means due to the survey design, it is tougher to get full credit on question 12 than on questions that accept only one answer such as question 6 and 37. The distribution of the students scores on question 12 is displayed in Figure 3. The question itself reads 'Which of the following are best practices in communicating with patients with IDD-MH and their families? Select all correct answers.' and two out of the four answer choices were correct. We would recommend that the training providers take a look at how they are covering the material that questions 6, 37, and 12 ask about (namely communication, common mental health conditions, and presentation of anxiety) and identify if there is a common theme in the method of presentation of these topics. If so, this could suggest that this method of instruction is highly effective for student retention and should be expanded to the other topics.

On the other hand, the results for question 22 and question 13, as displayed in Table 2, are concerning. Not only do these questions have the lowest amount of improvement in students accuracy after taking the training at +.1071 for question 22 and +.1887 for question 13, they also have the lowest average score on the post-training survey at .25 for question 22 and -.75 for question 13. This means that even after taking the full training, students are still selecting more incorrect than correct answers on question 13. Question 22 reads, 'Which is not a treatment approach that people diagnosed with intellectual and developmental disabilities and OCD may need?' and had one correct answer out of three answer choices. Question 13 reads 'If a patient with IDD-MH is acting out in the waiting room, you and your waiting room staff should: Select all correct answers' and had one correct answer out of 5 answer choices. As an outsider with minimal medical knowledge, these appear to be two critical questions that medical providers should know the answer to as treatment and waiting rooms are two common occurences in medicine. Thus, we would recommend that the training providers re-evaluate how they are covering treatment options and waiting room behavior as their current method is clearly ineffective. While students might enjoy how this material is being taught, the training is clearly not emphasizing the key points and conveying the true solution. This is a concerning result because medical students may falsely believe they have a stronger more accurate understanding of IDD patients after completing the training when they clearly do not.

Average Score on Questions Before and After Training

Question	Pre Survey	Post Survey	Change	Type
Q6	0.6734	1.0000	0.3265	One MC
Q22	0.1428	0.2500	0.1071	One MC
Q37	0.4489	0.8125	0.3635	One MC
Q10	0.2083	0.3611	0.1528	Hand Grade
Q11	0.6666	0.8055	0.1389	Hand Grade
Q30	0.2187	0.5000	0.2813	Hand Grade
Q10	0.3163	0.5833	0.2670	NLP Grade
Q11	0.5000	0.6388	0.1388	NLP Grade
Q30	0.2346	0.5555	0.3208	NLP Grade
Q12	0.5816	0.9375	0.3558	Multi MC
Q13	-0.9387	-0.7500	0.1887	Multi MC
Q23	0.2704	0.4843	0.2139	Multi MC
Q31	0.6122	0.8750	0.2627	Multi MC
Q34	0.3809	0.6458	0.2648	Multi MC
Q36	0.7040	0.9843	0.2802	Multi MC
Q39	0.3333	0.5625	0.2291	Multi MC

Table 2. This table displays the average scores of the respondents for all questions on the pre-training and post-training survey as well as the difference between these scores.

Next, we shift our analysis from the multiple choice questions to the short answer questions. We start with the results of our hand-grading. At baseline in the pre-assessment, participants scored the highest on the question, "How do you define culturally competent patient care?," with an average score of 0.67. For "How do you define positive medicine?" and "Why is the recognition of neurodiversity important in medical care?," participants scored lower at baseline, with average scores of 0.21 and 0.22, respectively.

The average score of all short answer questions increased in the post assessment, as visualized in Figure 4 and Figure 5. Participants continued to score the highest on the question about culturally competent care on the post-training survey, with the average score on that question being 0.8. The question about the importance of neurodiversity also saw improvement as it had an average score of 0.5 on the post-training survey compared to .22 on the pre-training survey. Participants scored the lowest on the question about positive medicine with an average score of 0.36. The score of 0.36 indicates that on average, participants were still not receiving partial credit for their responses even after taking the full training, as 0.5 points were awarded to individuals who answered part, but not all, of the question correctly according to our hand-grading methodology.

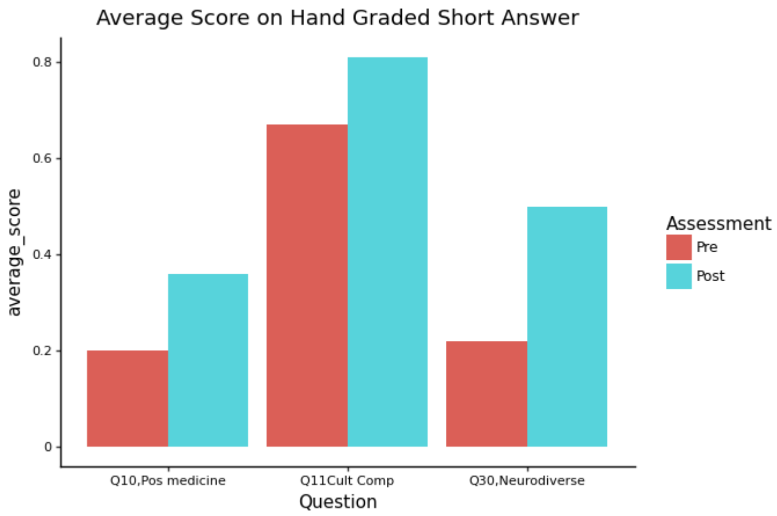

Fig. 4. The average scores of participants answering short answer questions 10, 11, and 30 on the pre and post surveys. Scores were determined via NLP analysis using the bag of words technique.

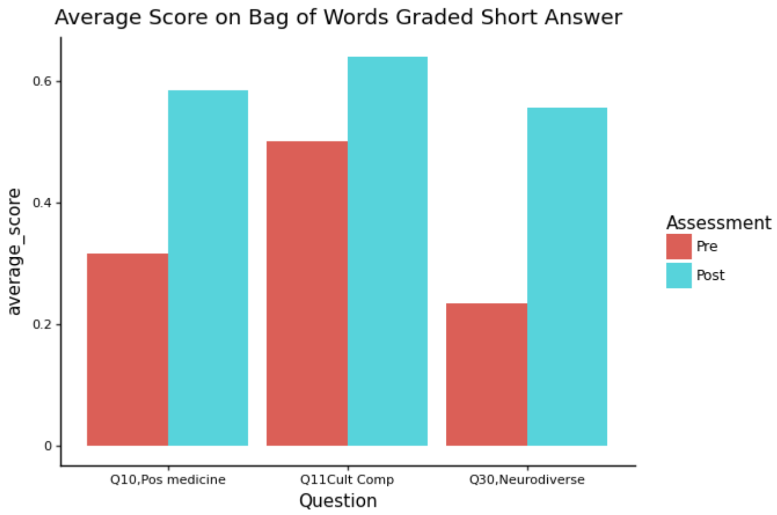

Fig. 5. The average scores of participants answering short answer questions 10, 11, and 30 on the pre and post surveys. Scores were determined via hand-grading.

The machine-graded questions score better on overall correctness with the post-assessment scores hovering around 0.6 for the three short answer questions. This means that the machine grader was most likely more attuned to select buzz words that a student used or it was receptive to a general topic area localized around the checked synonyms, even if the answer was not cohesive or rigorous enough to satisfy a human grader. Furthermore, while a similar proportion of correctness was maintained within a given question, the relative performance across the different questions was lost. On the human-graded questions, the question on positive medicine was understood significantly less compared to culturally competent care and neurodiversity. On the machine-graded questions, the positive medicine question looks like it performs equally as well as culturally competent care and neurodiversity.

While the question about culturally competent care had the highest average score on both the pre and post training survey, the percent difference between the pre and post training average scores for this question was the lowest out of all questions, at 13% according to both our hand grading and our NLP grading. On the other hand, the percent differences between the pre and post training surveys for "Why is the recognition of neurodiversity important in medical care?," was 28% when hand graded, and 32% when graded with NLP, indicating that students had the greatest amount of improvement on the question about the importance of neurodiversity. This should be viewed as an encouraging finding for the training developers as it identifies that while the medical students did initially lack the appropriate knowledge about the importance of neurodiversity, the training modules are proving effective in helping them learn and retain accurate information about neurodiversity.

Discussion. Overall, students who took part in the training program provided by The National Center for START Services saw improvement in their answers to questions about caring for IDD patients. As stated above, students had the worst post-training scores on questions 22 and 13, indicating that the training on treating IDD patients with OCD and communicating with IDD patients in the waiting room could have more improvement.

While this dataset was sufficient for preliminary analysis, there were a variety of limitations. First, there was a relatively small sample size and there was no way to track individual participants through each of the survey steps. Average scores across all participants were used to compensate for this, but it would likely be more helpful to track individual change for each question. This is specifically important because each of the students entered the pre-survey with a different amount of pre-existing knowledge. There is no way to know if the average increased because students who had no knowledge began to score higher, if the students that already had some knowledge were able to enhance their answers marginally and bring up the overall score, or if the lower scorers became frustrated by the material and dropped out of the training modules or neglected to complete the surveys. This also is an issue because multiple students dropped out at each step of the survey. As explained earlier, the pre-training survey had 49 eligible participants and the post-survey only had 16 respondents. There is no way to know if the average scores simply increased because the individuals who had scored poorly initially dropped out of the training.

In terms of grading the short answer questions, we chose to grade both by hand and NLP. While the output of the machine grader works well enough to analyze large trends across the cohort of respondents who took the IDD training modules, it loses nuance from the human-graded assessments and relies on an arbitrary threshold. This is one improvement which could be made upon this study: a better machine grader. This would also lift a limitation on this study which is scalability. The surveys were small and therefore manageable to grade by hand, but this might not be reproducible for larger studies in the future.

ACKNOWLEDGMENTS. We would like to thank the Dartmouth Center for Social Impact for connecting us with this project and allowing us to engage in meaningful work. Additionally, we would like to thank Professor Haber for providing us with the skills to conduct this work.

References.

1. C for Disease Control, Cdc's work on developmental disabilities. p. 1 (2022).
2. O Purugganan, Intellectual disabiilities. *Peds Rev* **39**, 299–309 (1994).