Revealing the Secret to Successful Virtual Meetings: How Personality, Social Skills, and More, Impact Conversational Involvement

Exploring the Dynamics of Conversational Involvement in Group Settings: The Influence of Individual Backgrounds

Ana Hobai

Supervisor(s): Catholijn Jonker, Masha Tsfasman

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Ana Hobai
Email of the student: A.Hobai@student.tudelft.nl
Final project course: CSE3000 Research Project
Thesis committee: Catholijn Jonker, Masha Tsfasman, Gosia Migut

Contents

1 INTRODUCTION **3**

2 BACKGROUND **3**

3 DATA **4**
3.1 Data pre-processing 4

4 METHODS **5**
4.1 Annotations 5
4.1.1 Definition of Group Involvement . . . 5
4.1.2 Annotation of Group Involvement . . 5
4.2 Questionnaire Data 5
4.3 Statistics 5
4.3.1 Data Exploration 5
4.3.2 Data Analysis 6
4.3.3 Data Modelling 6

5 RESULTS **6**

6 DISCUSSION **9**

7 CONCLUSION **9**
7.1 Summary 9
7.2 Limitations 9
7.2.1 Corpus Limitations 9
7.2.2 Annotations Limitations 10
7.2.3 Data Set Limitations 10
7.3 Future Work 10

8 RESPONSIBLE RESEARCH **10**

A Appendix **14**

Abstract

Since the recent rise and advancement of video conferencing platforms such as Zoom, it has become important to interpret the logistics of remote online meetings. Analysing verbal and non-verbal cues (such as body language) between members of these virtual forums can provide additional information regarding the level of conversational involvement of each party. This research focuses on age, gender, demographics and virtual background differences in the context of group conversational discussions. It argues that groups formed of younger adults have a higher level of involvement compared to the older groups. Similarly, this study found that groups with a higher ratio of male participants score better in virtual conversational engagement compared to women preponderant groups. To better understand the influence of these inter- and intra-personal characteristics, a corpus formed of 45 online meetings on the topic of Covid-19 was used. This set of data consists of questionnaires with measurements (demographics and other personal values), as well as detailed annotations of conversational signals, which provide valuable insights into the research topic of conversational involvement.

This study includes an experiment to investigate the involvement of individual backgrounds in the prediction of group conversational engagement. Four predictive models were used, namely the Decision Tree, Random Forest, Linear Regression and Generalized Linear Mixed Effects Models. While the Generalized Linear Mixed Effects Model provides more meaningful observations on the statistical effect of these factors, the Random Forest ultimately proved to give the best performance accuracy. The purpose of this research is to improve the connection between humans and technology by studying how inter- and intra- characteristics of individuals impact the involvement of a group in virtual interaction.

KEYWORDS: *group involvement, conversational involvement, individual backgrounds, age, gender, demographics, virtual experience, online meetings*

AKNOWLEDGEMENTS

I would like to show my sincere gratitude for the guidance of Catholijn Jonker and Masha Tsfasman throughout the process of conducting my research project, as well as thank Andy Li, Sebas de Melo e Silva Blasques de Holtreman, and Mathijs Rijm for their assistance in the early stages of this work.

1 INTRODUCTION

Effective communication is a fundamental aspect of human beings' social life, and group conversations provide a platform for individuals to exchange ideas, thoughts, and experiences. It is crucial to explore the dynamics of conversational involvement within group settings, particularly the influence of personal backgrounds, as they might provide valuable insights into group engagement. Individual backgrounds refer to the personal values, experiences, and characteristics of each group member.

Individual backgrounds, such as culture, gender, and age, can also affect communication within groups. Various studies showed that cultural and native language differences may lead to misconceptions [1], while age differences may result in power imbalances within the groups [2], [3]. Studying the impact of individual backgrounds on conversational involvement can provide insights into how communication can be enhanced within groups. For future reference, when 'engagement' is used, it is referred to as 'involvement' further in this study.

This research aims to show how personal backgrounds are of great importance when determining conversational involvement, which may impact the effectiveness of communication within groups. In the paper, current research on this topic will be examined, including studies and experiments conducted in recent years, as well as data exploration and analysis on a specific corpus.

By analysing previous literature on conversational involvement, this study seeks to develop a comprehensive understanding of what influences group communication. Having a deeper understanding of group conversational involvement will further help with the development of practical strategies for improving group interaction. This can also provide valuable insights for moderators and group managers.

The paper will begin by providing an overview of the concept of conversational involvement and its importance in group settings. It will then delve into the factors that influence conversational involvement, mainly the individual backgrounds. The data explored and analysed further in this paper will be referring to the MEMO corpus, which includes a set of online group discussions in the context of Covid-19.

Overall, this paper seeks to contribute to a better understanding of the dynamics of conversational involvement in group settings of virtual interactions, and how these dynamics can be influenced by personal backgrounds. The findings of this research could have implications in various fields, such as communication, psychology, and organizational behaviour.

2 BACKGROUND

Personal background factors, such as age, gender, ideology and native language differences, have been discovered to influence conversational involvement in group settings. For instance, previous research concluded men have lower involvement compared to women [4]. Individuals from diverse backgrounds can bring unique perspectives and experiences that can enhance group outcomes [5]. However, diversity can also lead to communication challenges, such as language barriers, cultural differences, and conflicting values, which can impede conversational involvement. Variations in the performance of native versus non-native individuals may produce misinterpretations and confusion [1]. A study on age stereotypes in the workplace has identified common stereotypes and potential moderators, emphasizing the need to consider age as an important factor when examining conversational involvement in intergenerational groups [3].

The research questions of this paper are the following:

Does the conversational involvement of a group change based on the individual backgrounds of each member?

(a) To what extent does age impact conversational involvement in a group setting?

(b) To what extent does gender influence the overall conversational engagement of a group?

(c) Do demographics and virtual meetings experience have any effect on group involvement in a virtual meeting?

These questions help build the central null hypothesis used in this study, *Individual backgrounds (namely age, gender, demographics and online discussions experience) have a significant impact on group conversational involvement*, as previous research also claims that age [6], gender [7], cultural differences [8] and other factors that form the individual backgrounds have various impacts on conversational involvement. For instance, young adults are more likely to have a clear speech, whereas the older adults class has more incentive to initiate conversations, which could lead to a moderate conversation in a group setting, as the generations have different perspectives and may be misinterpreted by the elderly individuals [9]. Based on this, it is expected to find that groups comprising a higher ratio of young adults would lead to an increased level of overall group involvement. Moreover, men are more engaged in face-to-face discussions, whereas women score a better involvement in online conversations [10], which leads to the hypothesis that groups with a higher percentage of females register better group engagement compared to the referenced gender. These previous studies have focused more on individual involvement concentrating only on students rather than any working class or age [10], or without the main scope of involvement discoveries, but rather to gain knowledge on the age class more inclined to speech disorders [9]. That is why this study aims to further pursue these findings. Concerning the cultural differences, an example between the Finnish and Japanese cultures states that silence may be more appropriate in some situations rather than an uninterrupted flow of speech [8]. Conversely, the main alternative hypothesis constructed as *Personal backgrounds do not have a significant impact on group engagement* will be proved false further in this paper, by making use of the age, gender and some background aspects (demographics and previous online experience) of the corpus participants to show that there is a correlation between the personal characteristics and the group conversational engagement.

By integrating findings from these studies, we can gain a comprehensive understanding of how individual backgrounds influence conversational involvement in group settings. These insights will contribute to our exploration of

the dynamics of conversational involvement and their implications for group interactions.

3 DATA

This research will be using the "MEMO corpus", which is a data collection corpus based on multimodal group discussions. This corpus consists of video recordings of virtual group meetings of 3–6 individuals over the course of three consecutive sessions, distanced 3–4 days apart. Each session lasts 45 minutes and includes a moderator who encourages interaction between participants and maintains the conversation flowing, containing discussions on the topic of Covid-19.

Overall, there were 53 MEMO corpus participants in total, consisting of 28 females, and 25 males, aged between 18 and 76 years old, as well as four moderators, three males, and one female, aged between 24 and 45 years old. All of them were fluent in English and resided in the United Kingdom. Each member has filled in a consent form for their participation in the experiment and agreed to the data collection using their signature. The members of this corpus were selected from various Covid-19 affected demographics: parents, students, business owners, and older adults (50+) in order to maximize the diversity of opinions in each group. To control the influence of previous relationships on group dynamics, participants and moderators met for the first time during their first meeting.

MEMO corpus contains around 34 hours of group interactions, 45 sessions, each of roughly 45 minutes, with a standard deviation of ±6.6 minutes [11]. Furthermore, each participant has filled in various surveys: a pre-screening survey, others before each session, and the post-screening after all sessions were completed. The questionnaires comprise a great variety of insightful variables and measures. However, the analysis used in this paper makes use of the demographics category selected from the pre-screening survey. The demographics section includes multiple variables such as age, gender, English fluency, country of residence and the Covid-19 affected group.

3.1 Data pre-processing

Previously selected variables (age, gender, virtual experience, and Covid-19 affected group as demographics) are preprocessed in such a way as to handle all missing values and inconsistent data. This case study excludes participants with missing values in any of the selected explanatory variables, as it can influence the level of representativeness of the samples, as well as introduce bias to the set [12]. The mean imputation of missing values has also been considered. However, a decision that it could damage the relationships among variables was reached [13]. Thus, this paper further makes use of the first method of handling missing values [14], as many other researchers also resort to removing instances which contain missing values [15], [16], [17].

The filtered data set comprises 35% duplicate entries, which are eliminated as they affect the quality of the data [18]. So, after the correct ID selection and duplicates and missing values in the fields of interest (age, gender, virtual experience or demographics) removal, the set is left with only 43 participants. Furthermore, the use of protected attributes, such as gender, age, race, marital status and others is not recommended as it may introduce bias in the analysis [19]. However, our research is based on the influence of these values to predict the overall group involvement, and as other studies show the preference of including these protected variables [20], this paper makes use of the following protected attributes: age, gender and demographics (Figure 1). These predictors are further analysed in the Results section to gain a deeper understanding of their correlation to group involvement.

age	GENDER	)emographi	online_meetings_experience	Group	nvolvemen
37	Female	parent	I've had online meetings before	3	3.19094
19	Female	business	I've had online meetings before	3	3.19094
59	Male	older	I've had online meetings before	3	3.19094
37	Female	business	I have online meetings on a regular basis	4	2.90384
20	Female	student	I have online meetings on a regular basis	4	2.90384

Figure 1: Values insights of the explanatory variables before encoding.

The independent variable, namely the group conversational engagement, is built on the annotations provided by the four different raters, each with distinct reasoning and perception. Thus, the first step in constructing the target variable is to comprehend how reliable the annotations set is. The author of this paper individually calculated the inter-annotator agreement by using three methods: Observed Agreement, Cohen's Kappa, and Intraclass Correlation (ICC). Observed Agreement is the division of the number of equal annotations and the overall number of annotations of the overlap between two raters, belonging to the interval [0.071, 0.428] in our case (Figure 28, in Appendix). Cohen's Kappa score (Figure 27, in Appendix) is frequently employed to evaluate the level of agreement between two annotators when subjects are categorized using a nominal scale, whereas the ICC is used on a numerical scale [21], aligning with our ordered dependent variable. Furthermore, the ICC type was determined based on a study which explains the ICC types [22], leading to the final choice of ICC3k (Figure 30 of the Appendix), calculated on the overlapping annotated segments between the raters (Figure 29, in Appendix). This score was chosen based on a 'Model', 'Type', and 'Definition' selection: the chosen model is the '2-way mixed effects model' since these four raters are the only annotators we use, the type is the k rater measurement, and the definition is 'consistency' as the ratings are cumulatively correlated [22].

As a result, the inter-annotator agreement scores range from 0.52 to 0.75, describing moderate reliability between the raters [23] (Figure 30, in Appendix). Nevertheless, it is important to include all annotations, as they represent different perspectives of the annotated task. The rather large variance between the four annotation sets may have been caused by the differences in personality and reasoning between raters [24], which will further be discussed in the Limitations chapter. The second annotator was more lenient, whereas the third rater was stricter when labelling the group involvement, leaving annotators one and four with more aligned annotation sets

(Figure 18, in Appendix). To handle the overlapping annotated segments, we used the mean between the ratings of the same moments. Previous studies show that it is hard to find the ground truth between multiple annotators [24], so the final target variable consists of all individual annotations and the means of the moments annotated by multiple raters.

4 METHODS

This section contains a thorough explanation of the methods used to answer the central question: 'Do individual backgrounds influence group conversational involvement?'. The first part of the research consisted of annotating the Corpus, helping us build our target variable, namely the group involvement variable from virtual group meetings. For this part, the ELAN software was used to annotate the videos of the multimodal corpus [25]. The second part consisted of exploring and analysing the data retrieved from the MEMO questionnaires, which will be used further in this paper to estimate and predict the target variable.

4.1 Annotations

Annotations undertake a central function in this research, as they provide insights and context to the data being analysed, more specifically, the involvement levels of groups in virtual interactions. The following subsections are meant as a guide through the stages of the annotation process. To start, the concepts of conversational involvement and group engagement were defined by analysing existing literature. Furthermore, extracting and labelling the group engagement from the provided corpus allows us to explore meaningful information on personal backgrounds of group conversational involvement in a virtual group meeting set-up.

4.1.1 Definition of Group Involvement

Successful communication creates an opportunity to develop and maintain group interactions through sharing various opinions, respecting conversational partners, accepting different perspectives, and creating relationships. For the rest of the research, it is important first to understand what conversational involvement means to be able to define group involvement as a whole. Conversational involvement on an individual basis refers to "the process by which individuals in an interaction start, maintain and end their perceived connection to one another" [26, p. 123]. Thus, we can now provide a complete definition of group involvement: "the perceived degree of interest or involvement of the majority of the group." [27, p. 490].

4.1.2 Annotation of Group Involvement

Group involvement has been annotated on a scale from 1 to 5, where 1 represents the lowest degree of group involvement, and 5 the highest. Each rater followed an annotation schema based on their own intuition by analysing both verbal (such as, but not limited to voice quality, intonation and verbal responses) and non-verbal (overall body language of each participant, such as eye gaze, facial expressions and body movements) cues to detect the overall group involvement. Each video in the corpus collection was split into random five seconds length segments which were also randomly assigned to four raters, each person having approximately six minutes of annotations per video to be done and around seventy-five annotations per video. The first group has not been annotated since the videos only had the person who was talking, not allowing us to accurately detect the group involvement level as the other meeting participants were not visible. Thus, there were about 3060 total annotations per person, each annotation having a five seconds length corresponding to a specific time frame in the video. The annotations include a 10 per cent overlap between each two annotators.

4.2 Questionnaire Data

The MEMO corpus comprises multiple Comma Separated Value (CSV) files containing the answers of participants to the questionnaires regarding demographics, personality, perception of the quality of other conversational partners and other factors which may influence the group conversational engagement. This collection of variables forms the dataset used in the modelling exercise, after being pre-processed. The pre-processed dataset is structured in several explanatory variables, which are used to estimate the defined target variable. After retrieving the data from all CSV files and merging it into one data set, it was further preprocessed to contain only relevant variables in this case study, such as demographics, experience with virtual meetings, gender, and age. The native language and ethnicity, variables relevant to our research, could not be analysed as all selected corpus participants were UK residents and fluent in English.

4.3 Statistics

This section is meant to give insight into the process of the three research steps of this study: Data Exploration, Data Analysis and Data Modelling. Data Exploration refers to the creation of the data set containing the explanatory and target variables, how it was built and why use the chosen methods. Data Analysis also consists of argumentation of the algorithms used to find the influence of the variables between each other. Lastly, the modelling section describes the prediction models used to estimate the group conversational involvement based on the explanatory variables selected in the preprocessing phase. The methods used in this experiment are discussed more in-depth in the following subsections.

4.3.1 Data Exploration

This section delves into interpreting the dataset used in this experiment, not only explanatory variables but also the target variable exploration. The independent variables previously mentioned and discussed in the Data section of this paper are represented by the age, gender, virtual experience and demographics of the corpus participants of each group. The target variable, group engagement, is represented by the annotations mentioned in the above sections.

When analysing the data forming the explanatory variables, it was noticed that some fields had the same entries, essentially representing the same class. For example, the perceived group and Covid-19 affected group variables were removed, as they held the same values as the demographics field. However, the choice to keep the demographics variable instead of the other two was made based on the number

of missing values in the Covid-19 affected group column and the names of the values, which had the same meaning (for example 'older', instead of 'Older adults (50+)') (Figure 9).

Furthermore, the demographics, gender and virtual experience classes are unordered categorical features, leading to the next step of the exploration analysis, the encoding of these non-numerical values. The label encoder uses integers to represent each value in the set, which might confuse the model into thinking that there is an actual order to the entries, even though there is no order to be considered [28]. Thus, due to the unordered nature of the categorical data, the One-Hot encoding method was used, since other investigations also make use of this measure [29]. This encoder turns each unique value from a field into a field of itself, comprising 0s when the participant does not belong to this group, and 1s when they do (Figure 10, Appendix).

4.3.2 Data Analysis

Data preparation is an essential step of this analysis, as it cleans the data set of unwanted entries [30]. Since the predictors have the same values for each session of each group, the target variable needed to be modified in such a way as to contain the mean involvement per group. This value was obtained as the average of the means of all three sessions. Even though the target variable was constructed on a scale from 1 to 5, with low to high involvement, the average involvement of each group would range between 2.75 and 3.35 (Figure 2), which was expected since conversations consist of both high and low moments for participation normally distributed throughout the meeting. Since the mean of the target variable would lead to a series of 3, the float value was kept instead for further modelling.

Categorical predictors have been one-hot encoded as previously mentioned, but there was still a need to check for multicollinearity between the variables. This experiment used heatmaps based on the correlation matrix between the predictors (Figures 14, 15b, 15a, in Appendix) to remove the 'older' class as this field overlapped with other columns, leading to multicollinearity. Age was kept over the older category as it provides more insightful observations. Also, the age variable was normalized to decrease the multicollinearity among predictors [31]. The regular virtual experience field was removed based on the Variance Inflation Factor (VIF) as it presented the highest score of 13.6787 among all variables (Figure 16a, in Appendix). Based on prior research [32], variables with a VIF score higher than 10 were eliminated. Thus, after removing the regular virtual experience variable, the VIF scores were calculated again to make sure there was no more collinearity between predictors (Figure 16b, in Appendix). The final data set comprised of these predictors (age, three classes of demographics, gender, and the virtual experience), and the overall involvement of each group.

4.3.3 Data Modelling

Modelling our data requires splitting our data into training and testing sets, in order not to overfit the model on the provided data set [33], as this would result in a perfect accuracy score, but could not be used to predict values when provided a new set of data to the model. This study uses four types of models to predict group involvement based on our explanatory variables: Linear Regression (LR), Decision Tree(DT), Random Forest(RF), and Generalized Mixed Model(GLMM). The multinomial regression was also considered, but as it needs the target variable to be unordered [34], and our group involvement has a hierarchy (1-low, 5-high) it could not be applied to our data set. The choice to use the first two predictive models was taken based on their complementary nature, providing useful information which cannot be obtained from the others [35]. Similarly, the use of the Generalized Mixed Model as a predictive model was based on previous studies that show how random effects variables, in this case, the group each participant belongs to, can influence the outcome of the normal linear regression if the dataset contains some kind of familiarity hierarchy [36]. Thus, it is useful to use the Mixed Effects Model as data is clustered in a specific way, namely, participants are clustered in groups [37]. The fixed variables in this model were the fields of interest when predicting group involvement: age, gender, demographics, and virtual experience, while the clustering (random effect) variable was the Group attribute since all participants belong to a specific group.

In the end, the K-Fold Cross-Validation method was used to split our data into training and testing and analyse the performance metrics for each model since the size of the sample data was small (the number of observations in each group). This type of performance assessment is also used to avoid overfitting data and evaluate the prediction error [38]. This method means splitting the data into k subsets and evaluating performance metrics using k-1 subsets as training and one as a validation set, but this process happens as many times as needed to have all subsets perform as validation. Then the average of all these errors, obtained from all k-validation sets, is computed as the accuracy error of the model [38]. This experiment used k=12, based on a study that explains how to choose the k-value [39, p. 5], in this case, k = $r * n$, where r = 0.3 (30% test set) and n is the total number of observations in the dataset (= 40).

5 RESULTS

Visualizing the variables provides a deeper understanding of what is wanted to be achieved. As expected, the group involvement mostly has scores of three (Figure 2), but it can be noticed a slight inclination towards better conversational involvement (higher than three). Even when analysing the involvement in each group, for example in group 5 session 1, the group involvement is moderate as we stated in our used annotation scale, but it still presents a subtle positive incline (Figure 11, in Appendix), as the number of (very-)high involvement is greater than the ones recorded as low (lower than three).

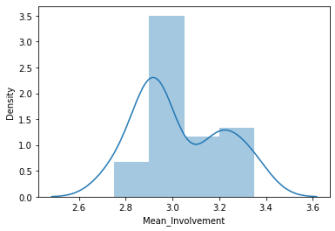


Figure 2: The overall involvement distribution.

Age is a numerical variable, so it was first individually analysed, to understand the relation to the target variable. Figure 3 shows that the corpus consists of an increased number of young adults, rather than the elderly class. However, there are other groups (Figure 12, in Appendix), such as group 15, where old adults dominate the age distribution, but this is not relevant since this group achieved moderate group involvement (Figure 5a).

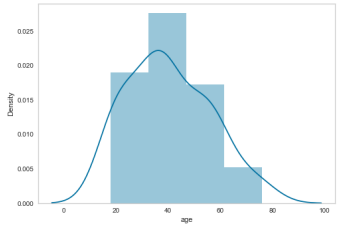


Figure 3: The age density over the entire dataset.

Moreover, the correlation between the age field and the group involvement is illustrated in Figure 4, describing a decrease in overall involvement when the elderly class participates in the virtual group meetings. Figure 5a also illustrates the influence of age on group involvement, as groups formed from more elderly people score a lower level of involvement compared to groups comprising teenagers and young adults. These results relate to the first research sub-question regarding the influence of age on overall group engagement, showing that as age increases, the target variable may be decreasing.

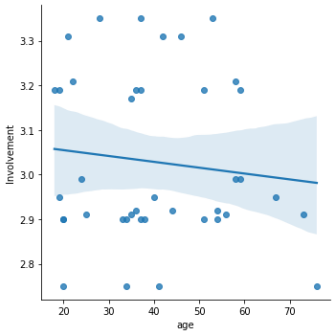
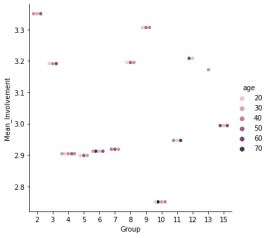
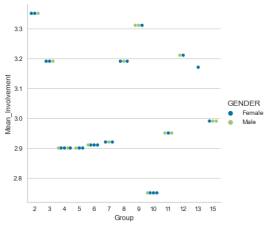


Figure 4: The correlation between age and involvement.



(a) Involvement of each group based on age.



(b) Involvement of each group based on gender.

Figure 5: Involvement plots based on (a) age and (b) gender

The gender of the participants is shown to be preponderant female (Figure 13, in Appendix). However, as shown in Figure 5b, the lowest levels of involvement were recorded in groups with a female ratio greater than the referenced gender, contradicting the null hypothesis on the gender difference in group involvement. Virtual experience and demographics were considered in this case study as complementary features of individual backgrounds, so they were analysed based on the main variables, age, gender and group involvement. Figure 6 depicts a pattern of involvement for each group based on the demographics of the group members. Group 2 has the highest involvement, consisting of mostly parents, business owners and older adults. Yet, Group 10 (lowest involvement) also registered the same demographics, but with a higher ratio of older adults. However, it is the only group with lower involvement than three to present almost the same demographics as the groups with higher registered involvement.

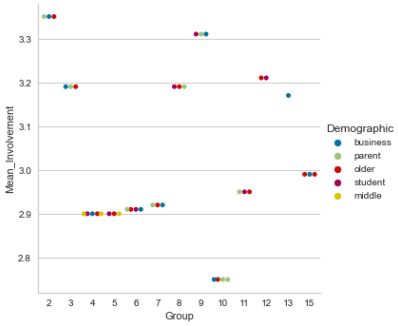


Figure 6: The relation involvement - demographics based on groups.

When analysing the relationship between gender and demographics concerning group involvement, it was found (Figure 17, in Appendix) that the groups with higher levels (higher than 3) of involvement contain more men than women. Regarding the virtual experience, Figure 7 clearly displays groups with participants who have had prior experience with online discussions scoring a better level of group involvement compared to the others. As the diversity of demographic and online backgrounds among participants of a group is higher, group engagement decreases, building an argument for the rest of the null hypothesis that demographics and previous online experience also influence group engagement.

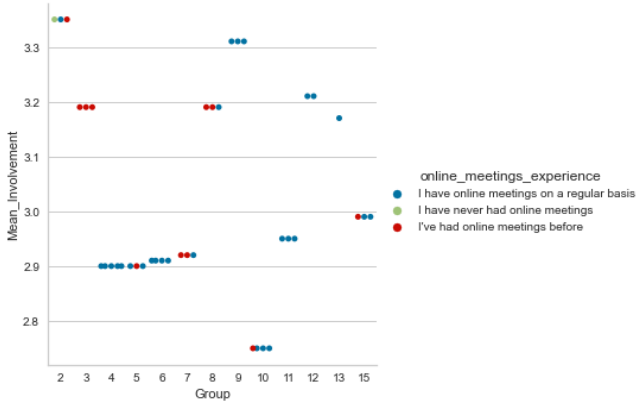


Figure 7: The involvement per group based on the virtual experience of participants.

Through the analysation of the dataset used for predictions, it has been found that no attribute is actually equally distributed (the same number of 0s and 1s). The age feature shows how the age of the participants ranges from 18 to 76 years old. Also, from Figure 19 of the Appendix, it is distinguishable that most participants are adults around the age of 40 since the mean of this variable is 40.5. The gender field

(Figure 20, in Appendix) has a mean of 0.4, proving once again our belief that the majority of participants are female (0 represents female, while 1 is for male). The next features are the demographic classes of each participant: the middle class represents the parents of young children (Figure 21, in Appendix), the mean of 0.08 shows that not many members belong to this class; the older demographic is represented by older adults (50+) (Figure 22, in Appendix), where the mean of 0.3 is higher than other classes, illustrating that most participants belong to this demographic category rather than the others; the parent field (Figure 23, in Appendix) presents a 22% of participants belonging to this class; the student (Figure 24, in Appendix) mean also suggests an 18% of the total participants as teenagers and young adults. The last predictor, the virtual_experience_Previous (Figure 25, in Appendix) exhibits that only 28% of the participants have had virtual meetings before this experiment, recording a standard deviation of 0.45 which displays evenly distributed data (the number of people who have had prior virtual meetings experience is almost equal to the number of people who either have virtual meetings regularly or have never participated in a virtual meeting). Finally, the dependent variable, the group involvement (Figure 26, in Appendix) has a mean of 3.03 which supports our previous statement that the overall group involvement is moderate-good.

Table 1: Performance metrics (k-fold cross-validation).

Model	RMSE	MAPE	MAE	MedAE
GLMM	0.324	0.093	0.286	0.285
Linear Regression	0.206	0.0606	0.185	0.181
Decision Tree	0.016	0.002	0.009	0.000
Random Forest	0.082	0.022	0.068	0.068

The best model performance was decided based on the Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), and Median Absolute Error (MedAE) metrics [40] [41]. Here is an explanation of how these values are obtained:

$$residuals = actual\_values - predicted\_values$$
$$absolute\_error = |actual\_values - predicted\_values|$$
$$RMSE = sqrt(mean((residuals)^2))$$
$$MAPE = mean(|residuals/actual\_values|) * 100$$
$$MAE = mean(absolute\_error)$$
$$MedAE = median(absolute\_error)$$

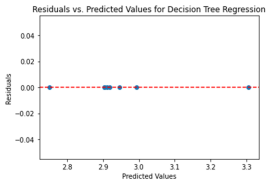
Table 1 displays the performance metrics of the prediction errors for these three models based on the k-fold cross-validation process. Based on the RMSE, MAPE, MAE, and MedAE metrics, models that scored lower values for these metrics are better predictors than the rest. The GLMM performs the worst with the highest results for all metrics, with a leading error of 0.32405 for RMSE, which translates to an RMSE accuracy of 67.595%. The LR Model was the second-worst predictive model, with the highest score of 0.206 for the RMSE. The Decision Tree (DT) Model performed the best, with an overall accuracy score of around 100% since the errors were between 0 (MedAE) and 0.005 (RMSE). However, the scores registered for the Random Forest Model were not far from the DT Model, with the highest error recorded for RMSE 0.089. The Random Forest Model consists of multiple decision trees, which create a better prediction and are less prone to overfitting data [42]. Thus, based on these results, the Random Forest Model performs the best among the four tested models, since the Decision Tree regressor seems to overfit the data.

The influence each predictor has on the target variable when using the Mixed Effects Model is described by their slopes (Table 2, in Appendix). The model converges, providing reliable estimates for the model parameters. The fixed effects of the predictors are indicated by the beta coefficients (Coef.), and the 95% confidence intervals indicate the range within the true value of the coefficient falls. In this case, all predictors have positive coefficients, which means that as these values increase, the group involvement also grows. This suggests that on average, male participants (based on the way data was encoded, male representing the value of 1) increase the level of involvement within a group, which answers the second research question concerning gender differences but contradicts the null hypothesis, stating that women are more active in virtual settings. Demographics and virtual experience also registered positive coefficients, having a positive impact on group engagement. The students variable recorded the highest coefficient among all demographics, meaning that groups with more students and young adults are more likely to have better conversational involvement. This aligns with the Data Analysis finding, where elders (60+) score lower involvement. In Figure 4, the number of young adults having an involvement lower than three is equal to the ones having a higher score, but the number of older adults (30+) who score lower is greater than the ones who score higher than three. Moreover, it was found through the slope of the previous virtual experience that as a group has more experience with online discussions, its involvement rises. These discoveries answer the research question regarding other factors from personal backgrounds influencing the target variable. Contradictory to Data Analysis and demographics results, the age coefficient of 0.69 presents higher engagement in groups with a higher ratio of older adults.

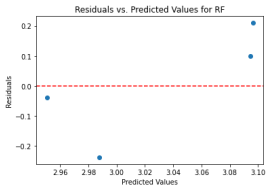
Furthermore, this experiment was conducted again with different target sets: the annotations set of the female rater and the set from the male raters (Tables 3, 4, in Appendix). The GLMM results from the female rater case show gender has a negative slope, meaning women increase group involvement. Whereas, the male raters found that men are more active in virtual group discussions.

Finally, the Residuals graphs show that the Random Forest Model gives the best predictions (Figure 8b). The Residuals graph for the Decision Tree Model predictions, shown in Figure 8a, illustrates that the model might actually overfit data since the residuals are all aligned to the y=0 axis, presenting a perfect fit of the model. The GLMM (Figure 8c) and the Linear Regression Model residuals graph (Figure 8d) are equally comparable since the residuals lie in the intervals [-0.3, 0.5] and [-0.5, 0.3], respectively, on the y-axis. These describe their poor performance since the residuals are the furthest from the y=0 axis, meaning the predictions made by

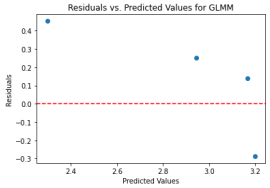
this model are very different from the actual values it was supposed to predict.



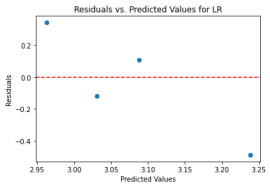
(a) Residuals graph for Decision Tree Model.



(b) Residuals graph for Random Forest Model.



(c) Residuals graph for Mixed Effect Model.



(d) Residuals graph for Linear Regression Model.

Figure 8: The residuals plots of the (a) Decision Tree, (b) Random Forest, (c) Mixed Effects, and (d) Linear Regression models.

6 DISCUSSION

Prior research showed age as a factor that needs to be considered when analysing the involvement of a group [3]. Groups formed of older adults achieve a higher level of engagement as a whole compared to the younger generation groups based on the GLMM results, as the age slope is positive, meaning that there is an incline in the group involvement when age grows. However, during Data Analysis, it has been found that elders (60+) score a lower involvement compared to younger generations (Figure 4). Elderly people are also more prone to speech impairment, which may be decreasing the level of concentration among participants, and reduce group engagement [12]. Thus, the null hypothesis derived from previous findings [12] concerning the age differences in discussions is assessed as true, relying on the results from the Data Analysis more since prior research also states that even if the model shows a good fit and data is consistent, there are still chances of random variation in the model [43].

Regarding gender, this research started from the hypothesis: men interact less in online discussions, and more in face-to-face conversations compared to women. This hypothesis contradicts the finding of this experiment, where men were found to be more active in online meetings from this corpus. This contradictory outcome may be based on the corpus used, as the topic of all online discussions was Covid-19, which impacted females from a more sensitive and emotional perspective compared to men, leaving them with an overwhelming feeling, stress, and even anxiety [44]. Despite the corpus comprising more female participants, this research found that the involvement of men was much higher. Also, the contradicted study on gender differences in online conversations [10] was conducted on a corpus containing a higher ratio of male participants than women, despite its finding of higher female involvement. Another reason for this controversial finding may relate to gender differences in the annotators. As shown in the results, from the female annotations, women are more involved, whereas, from the male raters' perspective, men increase the group engagement. Since there were three male raters and only one woman, the final results would be influenced by the male perspective.

Whether demographics and virtual experience impact group involvement is another aim of this research. The age finding aligns with the student demographics, as discovered to have a remarkable impact on the conversational involvement of a group, which also relates to the intergenerational connection study [6] arguing that younger adults are generally more involved in conversations compared to the elderly class. This may also be because younger people demonstrate clearer and more concise speech in conversations [9]. Moreover, online previous experience also recorded a positive incline in group involvement, proving that the online background of members also impacts the group engagement results.

7 CONCLUSION

7.1 Summary

Overall, this methodical research proves the null hypothesis that individual backgrounds influence the conversational involvement of a group, by conducting an experiment based on the Covid-19 topic-related multinomial corpus. The personal backgrounds this study focused on are age, gender, demographics and virtual experience. Groups consisting of younger people from the student demographic class were found to score an increased level of group involvement, proving the null hypothesis that older adults decrease group engagement, and answering the research question regarding the age effect, as well as the demographics impact, as the student demographic has a positive slope in relation to the group involvement. Moreover, the results answer the question concerning gender effects, as male preponderant groups were found to increase the overall conversational involvement. Thus, the null hypothesis: *Women are more involved in group conversations* is proved false. Previous experience with online discussions was also found as an influential personal characteristic when predicting group involvement, as groups with people who have had prior experience with virtual meetings score a better engagement to the others. Finally, this study discovered that group engagement can be best predicted by the Random Forest Model, showing that these personal characteristics provide enough information to further predict the group involvement.

7.2 Limitations

7.2.1 Corpus Limitations

This research focuses on the influence of personal backgrounds on group conversational engagement. However, the corpus used only focuses on the Covid-19 implications in the lives of participants, which may lead to topic-related biases. Moreover, the corpus used only consists of UK residents, therefore it may have an influence as language can rep-

resent a barrier too, which should be taken into account when analysing individual backgrounds.

7.2.2 Annotations Limitations

Raters did not have previous experience with the annotation process, thus, it is hard to depict which annotations represent the ground truth. Likewise, raters did not receive professional training before starting to annotate the corpus. As a result, the novice level of annotators resulted in a moderate inter-annotator agreement score [45]. However, novice annotators were shown to have a performance comparable to the usage of annotations provided by expert raters [46]. Furthermore, the number of annotators is another factor which may have influenced the score of the inter-rater reliability, as the score would be significantly higher if the study would have had only two annotators instead of four [45].

7.2.3 Data Set Limitations

This experiment also holds its own limitations concerning the data set particularly. Firstly, the questionnaires contained some missing values, meaning that some participants did not provide the full overview of their backgrounds, which led to these participants being omitted from this research, impacting the actual relation analysis between individual backgrounds and the group engagement score. Also, some prolific IDs were written incorrectly, so only those IDs that matched the ones from the main CSV were used, reducing the number of observations drastically. Secondly, the target set was represented by the annotations done as a first step in this study, so all limitations from the annotations apply here as well. Since there were noticeable differences between annotators, the target variable may be influenced by the instinct of each rater. There were multiple options when considering how to combine the four annotation sets of each rater, such as eliminating the two sets that contrasted each other the most and taking the sets from the two annotators who were more aligned as the ground truth. However, the study wants to include all annotation sets, as all raters have different reasoning, so all sets provided by the annotators who were both lenient and strict were included.

7.3 Future Work

This research could still be enhanced by thoroughly analysing the models used, as well as any information received, such as the errors, which could be differently investigated and consulted with experts. Furthermore, an Optimization Analysis of the models used should also be conducted, including hyperparameters tuning, to achieve the best prediction results. Besides this, the demographics variable should be considered more carefully, as it may give more insightful information when analysed individually, rather than in relation to other predictors.

8 RESPONSIBLE RESEARCH

When conducting this study, various ethics need to be considered. Such ethical considerations help keep privacy control of confidential data, support the credibility of the results, and easily reproduce the research when the person conducting the replication has access to the corpus. Firstly, this research makes use of data provided by the confidential MEMO corpus. Therefore, only people with access to this corpus may be able to replicate the work that was carried out in this study. Due to confidential reasons, this study can't reveal any personal information of any participant used when constructing the data set, or any pictures of the participants when explaining the guidelines of the annotating scheme. Secondly, the paper introduces both previously used claims, findings and methods, as well as new discoveries. Already researched assertions and results are supported by proper references when introduced in this research. Decisions of methods to use and claims to make were taken after exploring multiple studies in the specific field based on the keywords of interest. Thirdly, the experiment can be easily reproduced if the person replicating it has access to the MEMO Corpus as the source code for Data Analysis and Modelling has been made public on GitHub [47].

References

[1] B. Wrede and E. Shriberg, "Spotting "hot spots" in meetings: Human judgments and prosodic cues," in *8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, ISCA, Sep. 2003. DOI: 10.21437/eurospeech.2003-747. [Online]. Available: https://doi.org/10.21437/eurospeech.2003-747.

[2] A. Burmeister, A. Hirschi, and H. Zacher, "Explaining Age Differences in the Motivating Potential of Intergenerational Contact at Work," *Work, Aging and Retirement*, vol. 7, no. 3, pp. 197–213, Apr. 2021, ISSN: 2054-4650. DOI: 10.1093/workar/waab002. eprint: https://academic.oup.com/workar/article-pdf/7/3/197/38816704/waab002.pdf. [Online]. Available: https://doi.org/10.1093/workar/waab002.

[3] R. Posthuma and M. Campion, "Age stereotypes in the workplace: Common stereotypes, moderators, and future research directions†," *Journal of Management - J MANAGE*, vol. 35, pp. 158–188, Feb. 2009. DOI: 10.1177/0149206308318617.

[4] H. E. Krugman, "The measurement of advertising involvement," *The Public Opinion Quarterly*, vol. 30, no. 4, pp. 583–596, 1966, ISSN: 0033362X, 15375331. [Online]. Available: http://www.jstor.org/stable/2746964 (visited on 05/30/2023).

[5] J. Hollenbeck, B. Beersma, and M. Schouten, "Beyond team types and taxonomies: A dimensional scaling conceptualization for team description," *Academy of Management Review*, vol. 37, pp. 82–106, Jan. 2012. DOI: 10.5465/amr.2010.0181.

[6] A. S. Brown, E. M. Jones, and T. L. Davis, "Age differences in conversational source monitoring.," *Psychology and Aging*, vol. 10, no. 1, p. 111, 1995.

[7] B. Sun, H. Mao, and C. Yin, "Male and female users' differences in online technology community based on text mining," *Frontiers in Psychology*, vol. 11, p. 806, May 2020. DOI: 10.3389/fpsyg.2020.00806.

[8] N. Koudenburg, T. Postmes, and E. H. Gordijn, "Beyond content of conversation: The role of conversational form in the emergence and regulation of social structure," *Personality and Social Psychology Review*, vol. 21, no. 1, pp. 50–71, 2017.

[9] N. Pereira, A. P. B. Gonçalves, M. Goulart, M. A. Tarrasconi, R. Kochhann, and R. P. Fonseca, "Age-related differences in conversational discourse abilities a comparative study," *Dementia & Neuropsychologia*, vol. 13, no. 1, pp. 53–71, Jan. 2019, ISSN: 1980-5764. DOI: 10.1590/1980-57642018dn13-010006. [Online]. Available: https://doi.org/10.1590/1980-57642018dn13-010006.

[10] M.-J. Tsai, J.-C. Liang, H.-T. Hou, and C.-C. Tsai, "Males are not as active as females in online discussion: Gender differences in face-to-face and online discussion strategies," *Australasian Journal of Educational Technology*, vol. 2015, pp. 263–277, May 2015. DOI: 10.14742/ajet.1557.

[11] M. Tsfasman, K. Fenech, M. Tarvirdians, A. Lorincz, C. Jonker, and C. Oertel, "Towards creating a conversational memory for long-term meeting support: Predicting memorable moments in multi-party conversations through eye-gaze," in *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION*, ACM, Nov. 2022. DOI: 10.1145/3536221.3556613. [Online]. Available: https://doi.org/10.1145/3536221.3556613.

[12] H. Kang, "The prevention and handling of the missing data," *Korean journal of anesthesiology*, vol. 64, pp. 402–6, May 2013. DOI: 10.4097/kjae.2013.64.5.402.

[13] C. Glas, "Missing data," in *International Encyclopedia of Education (Third Edition)*, P. Peterson, E. Baker, and B. McGaw, Eds., Third Edition, Oxford: Elsevier, 2010, pp. 283–288, ISBN: 978-0-08-044894-7. DOI: https://doi.org/10.1016/B978-0-08-044894-7.01346-4. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780080448947013464.

[14] M. Jamshidian and M. Mata, "2 - advances in analysis of mean and covariance structure when data are incomplete**this research was supported in part by the national science foundation grant dms-0437258.," in *Handbook of Latent Variable and Related Models*, ser. Handbook of Computing and Statistics with Applications, S.-Y. Lee, Ed., Amsterdam: North-Holland, 2007, pp. 21–44. DOI: https://doi.org/10.1016/B978-044452044-9/50005-7. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780444520449500057.

[15] E. Choi, Z. Xu, Y. Li, *et al.*, "Proceedings of the aaai conference on artificial intelligence.," 2020.

[16] V. Iosifidis and E. Ntoutsi, "Dealing with bias via data augmentation in supervised learning scenarios," 2018.

[17] V. Iosifidis and E. Ntoutsi, "Adafair: Cumulative fairness adaptive boosting," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, ser. CIKM '19, Beijing, China: Association for Computing Machinery, 2019, pp. 781–790, ISBN: 9781450369763. DOI: 10.1145/3357384.3357974. [Online]. Available: https://doi-org.tudelft.idm.oclc.org/10.1145/3357384.3357974.

[18] W. Lup Low, M. Li Lee, and T. Wang Ling, "A knowledge-based approach for duplicate elimination in data cleaning," *Information Systems*, vol. 26, no. 8, pp. 585–606, 2001, Data Extraction,Cleaning and Reconciliation, ISSN: 0306-4379. DOI: https://doi.org/10.1016/S0306-4379(01)00041-2. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306437901000412.

[19] A. Romanov, M. De-Arteaga, H. Wallach, *et al.*, "What's in a name? reducing bias in bios without access to protected attributes," *arXiv preprint arXiv:1904.05233*, 2019.

[20] R. Yu, H. Lee, and R. F. Kizilcec, "Should college dropout prediction models include protected at-

tributes?" In *Proceedings of the Eighth ACM Conference on Learning @ Scale*, ser. L@S '21, Virtual Event, Germany: Association for Computing Machinery, 2021, pp. 91–100, ISBN: 9781450382151. DOI: 10.1145/3430895.3460139. [Online]. Available: https://doi-org.tudelft.idm.oclc.org/10.1145/3430895.3460139.

[21] M. J. Warrens, "Five ways to look at cohen's kappa," *Journal of Psychology & Psychotherapy*, vol. 5, no. 4, p. 1, 2015.

[22] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155–163, 2016, ISSN: 1556-3707. DOI: https://doi.org/10.1016/j.jcm.2016.02.012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1556370716000158.

[23] C. A. Bobak, P. J. Barr, and A. J. O'Malley, "Estimation of an inter-rater intra-class correlation coefficient that overcomes common assumption violations in the assessment of health measurement scales," *BMC Medical Research Methodology*, vol. 18, no. 1, Sep. 2018. DOI: 10.1186/s12874-018-0550-6. [Online]. Available: https://doi.org/10.1186/s12874-018-0550-6.

[24] Y. Yan, R. Rosales, G. Fung, R. Subramanian, and J. Dy, "Learning from multiple annotators with varying expertise," *Machine Learning*, vol. 95, Jun. 2014. DOI: 10.1007/s10994-013-5412-1.

[25] *ELAN (Version 6.5) [Computer software]*, en, https://archive.mpi.nl/tla/elan, (2023). Nijmegen: Max Planck Institute for Psycholinguistics.

[26] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots," *Artificial Intelligence*, vol. 166, no. 1-2, pp. 140–164, Aug. 2005. DOI: 10.1016/j.artint.2005.03.005. [Online]. Available: https://doi.org/10.1016/j.artint.2005.03.005.

[27] D. Gatica-Perez, I. McCowan, D. Zhang, and S. Bengio, "Detecting group interest-level in meetings," in *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, IEEE. DOI: 10.1109/icassp.2005.1415157. [Online]. Available: https://doi.org/10.1109/icassp.2005.1415157.

[28] E. Jackson and R. Agrawal, *Performance evaluation of different feature encoding schemes on cybersecurity logs*. IEEE, 2019.

[29] M. Alruily, S. A. El-Ghany, A. M. Mostafa, M. Ezz, and A. A. A. El-Aziz, "A-tuning ensemble machine learning technique for cerebral stroke prediction," *Applied Sciences*, vol. 13, no. 8, 2023, ISSN: 2076-3417. DOI: 10.3390/app13085047. [Online]. Available: https://www.mdpi.com/2076-3417/13/8/5047.

[30] V. Kotu and B. Deshpande, "Chapter 3 - data exploration," in *Data Science (Second Edition)*, V. Kotu and B. Deshpande, Eds., Second Edition, Morgan Kaufmann, 2019, pp. 39–64, ISBN: 978-0-12-814761-0. DOI: https://doi.org/10.1016/B978-0-12-814761-0.00003-4. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128147610000034.

[31] D. W. Marquardt, "Comment: You should standardize the predictor variables in your regression models," *Journal of the American Statistical Association*, vol. 75, no. 369, pp. 87–91, 1980.

[32] G. R. Franke, "Multicollinearity," *Wiley international encyclopedia of marketing*, 2010.

[33] Y. Bai, M. Chen, P. Zhou, *et al.*, "How important is the train-validation split in meta-learning?" In *International Conference on Machine Learning*, PMLR, 2021, pp. 543–553.

[34] C. Kwak and A. Clayton-Matthews, "Multinomial logistic regression," *Nursing research*, vol. 51, no. 6, pp. 404–410, 2002.

[35] W. J. Long, J. L. Griffith, H. P. Selker, and R. B. D'agostino, "A comparison of logistic regression to decision-tree induction in a medical domain," *Computers and Biomedical Research*, vol. 26, no. 1, pp. 74–97, 1993.

[36] B. M. Bolker, "Linear and generalized linear mixed models," *Ecological statistics: contemporary theory and application*, pp. 309–333, 2015.

[37] S. Rabe-Hesketh and A. Skrondal, "Generalized linear mixed models," in *International Encyclopedia of Education (Third Edition)*, P. Peterson, E. Baker, and B. McGaw, Eds., Third Edition, Oxford: Elsevier, 2010, pp. 171–177, ISBN: 978-0-08-044894-7. DOI: https://doi.org/10.1016/B978-0-08-044894-7.01332-4. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780080448947013324.

[38] D. Berrar, *Cross-validation.* 2019.

[39] Z. Mahmood and S. Khan, "On the use of k-fold cross-validation to choose cutoff values and assess the performance of predictive models in stepwise regression," *The International Journal of Biostatistics*, vol. 5, no. 1, 2009.

[40] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance," *Climate research*, vol. 30, no. 1, pp. 79–82, 2005.

[41] T. Chai and R. R. Draxler, "Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature," *Geoscientific model development*, vol. 7, no. 3, pp. 1247–1250, 2014.

[42] T. Lan, H. Hu, C. Jiang, G. Yang, and Z. Zhao, "A comparative study of decision tree, random forest, and convolutional neural network for spread-f identification," *Advances in Space Research*, vol. 65, no. 8, pp. 2052–2061, 2020.

[43] Z. Pan and D. Lin, "Goodness-of-fit methods for generalized linear mixed models," *Biometrics*, vol. 61, no. 4, pp. 1000–1009, 2005.

[44] S. Hennekam and Y. Shymko, "Coping with the covid-19 crisis: Force majeure and gender performativity," *Gender, Work & Organization*, vol. 27, no. 5, pp. 788–803, 2020.

[45] P. S. Bayerl and K. I. Paul, "What Determines Inter-Coder Agreement in Manual Annotations? A Meta-Analytic Investigation," *Computational Linguistics*, vol. 37, no. 4, pp. 699–725, Dec. 2011, ISSN: 0891-2017. DOI: 10.1162/COLI_a_00074. eprint: https://direct.mit.edu/coli/article-pdf/37/4/699/1798897/coli_a_00074.pdf. [Online]. Available: https://doi.org/10.1162/COLI%5C_a%5C_00074.

[46] S. Budd, T. Day, J. Simpson, *et al.*, "Can non-specialists provide high quality gold standard labels in challenging modalities?" In *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health: Third MICCAI Workshop, DART 2021, and First MICCAI Workshop, FAIR 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27 and October 1, 2021, Proceedings 3*, Springer, 2021, pp. 251–262.

[47] A. Hobai, *Machine Learning Project*, 2023. [Online]. Available: https://github.com/ahobai/ResearchProject.git.

A Appendix

rofilicIt	Group	english_fluency	country_of_residence	vid-19_affected_grc	Demographic	Perceived_group	
60fe…	3	Advanced	United Kingdom	nan	business	business	I've
5b70…	3	Native	UK	nan	older	older	I've
613a…	4	Native	UK	nan	business	business	I ha
6139…	4	Native	united kingdom	Students	student	Students	I ha
5d03…	4	Native	UK	Older adults …	older	Older adults (50…	I ha
6128…	4	Advanced	United Kingdom	Parents of young children	middle	Parents of young children	I ha
5ea7…	5	Advanced	uk	Older adults (50+)	older	Older adults (50+)	I've
6126…	5	Native	United Kingdom	Students	student	Students	I ha
613a…	5	Advanced	UK	Parents of young children	middle	Parents of young children	I ha
613a…	4	Advanced	Scotland	Parents of yo…	middle	Parents of young…	I ha
613b…	2	Advanced	United Kingdom	nan	business	business	I ha

Figure 9: Demographics and perceived group differences

Index	age	GENDER_Male	middle	older	parent	student	virtual_experienc_Regular	virtual_experience_Previous
0	37	False	False	False	True	False	False	True
1	19	False	False	False	False	False	False	True
2	59	True	False	True	False	False	False	True
3	37	False	False	False	False	False	True	False
4	20	False	False	False	False	True	True	False
5	51	True	False	True	False	False	True	False
6	34	True	True	False	False	False	True	False
7	38	False	True	False	False	False	True	False
8	54	False	False	True	False	False	False	True
9	20	True	False	False	False	True	True	False
10	33	False	True	False	False	False	True	False

Figure 10: Preview after encoding of categorical data.

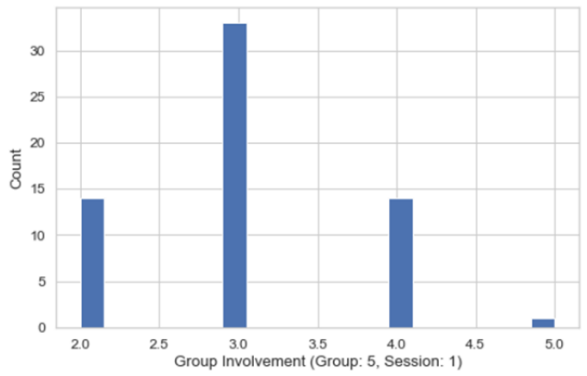


Figure 11: The involvement distribution in group 5, session 1

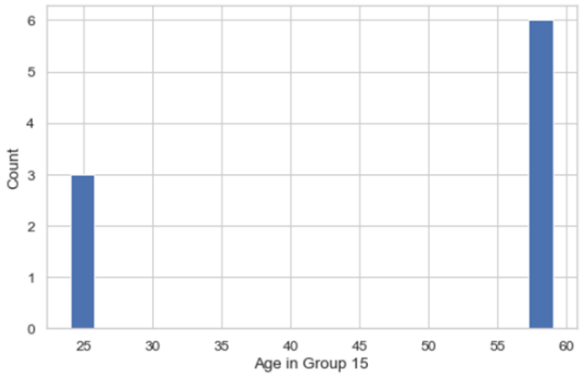


Figure 12: Age distribution in group 15.

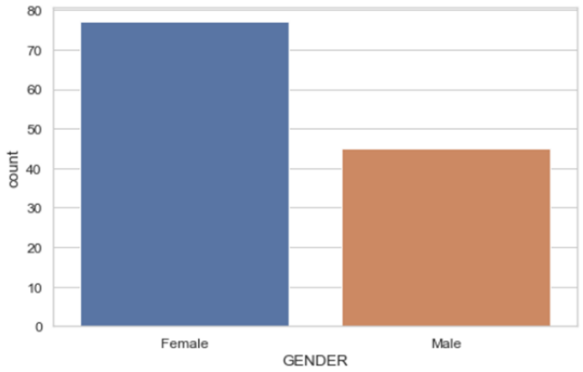


Figure 13: Overall gender distribution.

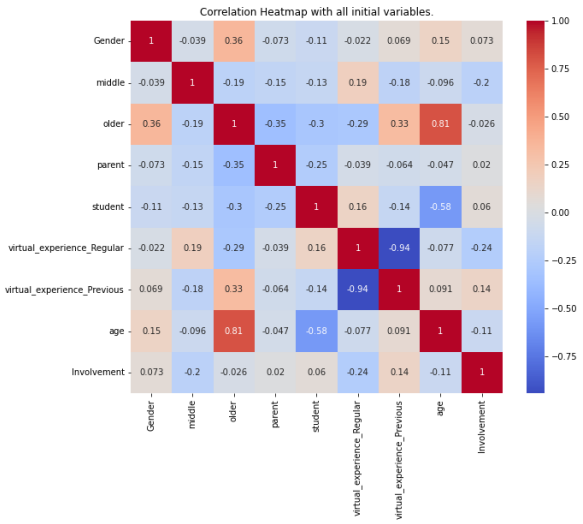
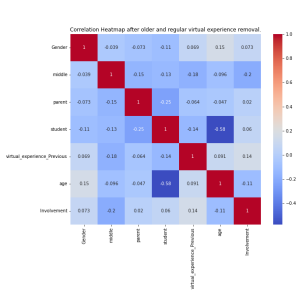
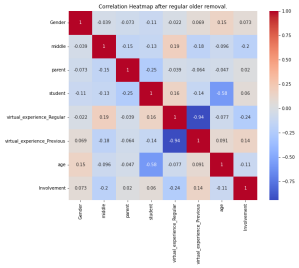


Figure 14: The heatmap of the correlation matrix between all initial predictors.



(a) Removal of virtual_experience_Regular.



(b) Removal of the older category as it intersects with the age field.

Figure 15: The heatmap of the correlation matrix between all variables in the data set after removal of (a) virtual_experience_Regular and (b) older fields.

Variable	VIF
age	10.2022
Gender	1.77118
middle	1.36198
parent	1.35378
student	2.01179
virtual_experience_Regular	13.6787
virtual_experience_Previous	4.90319
Group	6.42613

(a) After the heatmap reduction (removal of older).

Variable	VIF
age	5.09775
Gender	1.75332
middle	1.09756
parent	1.28624
student	1.31764
virtual_experience_Previous	1.42276
Group	4.88989

(b) After the 'virtual_experience_Regular' category removal.

Figure 16: The VIF scores of the explanatory variable.

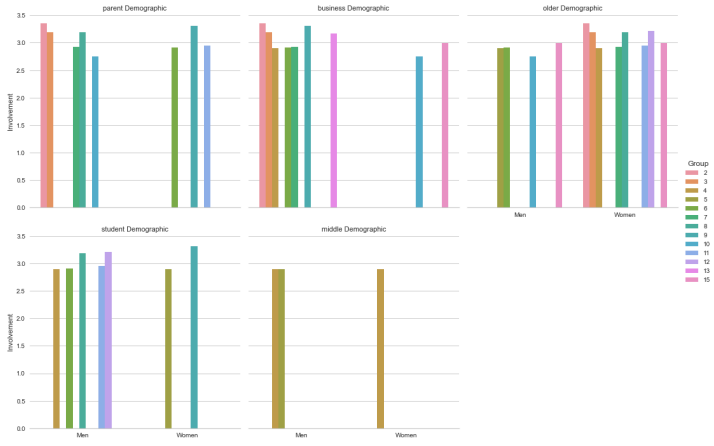


Figure 17: The involvement per group based on gender and demographics.

Group	Annotator1	Annotator2	Annotator3	Annotator4
2	3	4	2	3
3	3	4	2	3
4	3	4	2	3
5	3	3	2	3
6	3	3	2	3

Figure 18: The mean overall involvement per group from each annotator for groups 2-6.

```
Variable: age    Data Type: float64

 Number of Obs.: 40
 Number of missing obs.: 0
 Percent missing: 0.0
 Number of unique values: 28

 Range: [18.0, 76.0]
 Mean: 40.25
 Standard Deviation: 15.78
```

Figure 19: Age attribute analysis.

```
Variable: Gender    Data Type: float64

 Number of Obs.: 40
 Number of missing obs.: 0
 Percent missing: 0.0
 Number of unique values: 2

 Range: [0.0, 1.0]
 Mean: 0.4
 Standard Deviation: 0.5
```

Figure 20: Gender attribute analysis.

```
Variable: middle    Data Type: float64

 Number of Obs.: 40
 Number of missing obs.: 0
 Percent missing: 0.0
 Number of unique values: 2

 Range: [0.0, 1.0]
 Mean: 0.08
 Standard Deviation: 0.27
```

Figure 21: Demographics, middle (Parents of young children) analysis.

```
Variable: older    Data Type: float64

 Number of Obs.: 40
 Number of missing obs.: 0
 Percent missing: 0.0
 Number of unique values: 2

 Range: [0.0, 1.0]
 Mean: 0.3
 Standard Deviation: 0.46
```

Figure 22: Demographics, older category analysis.

```
Variable: parent    Data Type: float64

 Number of Obs.: 40
 Number of missing obs.: 0
 Percent missing: 0.0
 Number of unique values: 2

 Range: [0.0, 1.0]
 Mean: 0.22
 Standard Deviation: 0.42
```

Figure 23: Demographics, parent category analysis.

```
Variable: student    Data Type: float64

 Number of Obs.: 40
 Number of missing obs.: 0
 Percent missing: 0.0
 Number of unique values: 2

 Range: [0.0, 1.0]
 Mean: 0.18
 Standard Deviation: 0.38
```

Figure 24: Demographics, student category analysis.

```
Variable: virtual_experience_Previous    Data Type: float64

 Number of Obs.: 40
 Number of missing obs.: 0
 Percent missing: 0.0
 Number of unique values: 2

 Range: [0.0, 1.0]
 Mean: 0.28
 Standard Deviation: 0.45
```

Figure 25: Virtual experience analysis.

```
Variable: Mean_Involvement    Data Type: float64

 Number of Obs.: 40
 Number of missing obs.: 0
 Percent missing: 0.0
 Number of unique values: 13

 Range: [2.7504009892245187, 3.3501766472354704]
 Mean: 3.03
 Standard Deviation: 0.19
```

Figure 26: Group involvement analysis.

Table 2: The random intercepts (z-statistics, corresponding p-values), random slopes (coef), and random variance components (Std. Err.) obtained from fitting data into the Linear Mixed Effects Model.

Variable	Coef.	Std. Err.	[0.025	0.975]
age	0.690	0.050	0.592	0.787
Gender	0.085	0.075	-0.063	0.232
middle	0.285	0.153	-0.014	0.584
parent	0.276	0.094	0.091	0.460
student	0.707	0.103	0.505	0.909
virtual_experience_Previous	0.135	0.106	-0.072	0.343
Group Var	0.033	0.167		

Table 3: The random slopes (coef), and random variance components (Std. Err.) obtained from fitting data into the Linear Mixed Effects Model when using only the annotations from the female rater.

Variable	Coef.	Std.Err.	[0.025	0.975]
age	0.892	0.064	0.766	1.018
Gender	-0.011	0.122	-0.251	0.228
middle	0.265	0.242	-0.209	0.738
parent	0.200	0.145	-0.085	0.485
student	0.786	0.159	0.473	1.098
virtual_experience_Previous	0.182	0.161	-0.134	0.498
Group Var	0.035	0.152		

Table 4: The random slopes (coef), and random variance components (Std. Err.) obtained from fitting data into the Linear Mixed Effects Model when using only the annotations from the male raters.

Variable	Coef.	Std.Err.	[0.025	0.975]
age	0.689	0.053	0.585	0.792
Gender	0.018	0.094	-0.166	0.202
middle	0.188	0.190	-0.184	0.559
parent	0.159	0.112	-0.060	0.378
student	0.650	0.127	0.402	0.898
virtual_experience_Previous	0.076	0.121	-0.161	0.312
Group Var	0.028	0.132		

kappa12	float64	1	0.07982438635002975
kappa13	float64	1	-0.017202036159382095
kappa14	float64	1	0.07830554749389995
kappa23	float64	1	-0.05285377667233937
kappa24	float64	1	0.014421445537834976
kappa34	float64	1	0.11677911250068573

Figure 27: The kappa score for inter-rater agreement

observed_agreement12	float	1	0.07142857142857142
observed_agreement13	float	1	0.2857142857142857
observed_agreement14	float	1	0.42857142857142855
observed_agreement23	float	1	0.07142857142857142
observed_agreement24	float	1	0.42857142857142855
observed_agreement34	float	1	0.14285714285714285

Figure 28: Observed agreement between the raters

Index	Group_session_time	Annotator	nvolvemen
8	g2-s1-t490000	1	3
13	g2-s1-t685000	1	4
29	g2-s1-t1365000	1	2
34	g2-s1-t1540000	1	2
40	g2-s1-t1775000	1	2
48	g2-s1-t2145000	1	3
55	g2-s1-t2370000	1	3
59	g2-s1-t2430000	1	3
61	g2-s1-t2465000	1	2
76	g2-s2-t490000	1	4
81	g2-s2-t685000	1	1

Index	Group_session_time	Annotator	nvolvemen
2636	g2-s1-t490000	2	5
2647	g2-s1-t685000	2	5
2694	g2-s1-t1365000	2	3
2706	g2-s1-t1540000	2	3
2723	g2-s1-t1775000	2	3
2747	g2-s1-t2145000	2	4
2764	g2-s1-t2370000	2	4
2767	g2-s1-t2430000	2	4
2769	g2-s1-t2465000	2	3
2809	g2-s2-t490000	2	5
2820	g2-s2-t685000	2	3

Figure 29: The overlapping annotated segments between rater 1 and rater 2.

Type	Description	ICC	F	df1	df2	pval	CI95%
ICC1	Single raters absolute	0.152458	1.35977	173	174	0.0218207	[0. 0.29]
ICC2	Single random raters	0.284249	2.40455	173	173	7.00982e-09	[-0.01 0.51]
ICC3	Single fixed raters	0.412551	2.40455	173	173	7.00982e-09	[0.28 0.53]
ICC1k	Average raters absolute	0.264579	1.35977	173	174	0.0218207	[0.01 0.45]
ICC2k	Average random raters	0.44267	2.40455	173	173	7.00982e-09	[-0.02 0.67]
ICC3k	Average fixed raters	0.584122	2.40455	173	173	7.00982e-09	[0.44 0.69]

(a) ICC between rater 1 and rater 2

Type	Description	ICC	F	df1	df2	pval	CI95%
ICC1	Single raters absolute	0.181272	1.44281	121	122	0.0221404	[0. 0.35]
ICC2	Single random raters	0.354052	4.16291	121	121	2.53293e-14	[-0.1 0.66]
ICC3	Single fixed raters	0.612622	4.16291	121	121	2.53293e-14	[0.49 0.71]
ICC1k	Average raters absolute	0.30691	1.44281	121	122	0.0221404	[0.01 0.52]
ICC2k	Average random raters	0.522952	4.16291	121	121	2.53293e-14	[-0.21 0.79]
ICC3k	Average fixed raters	0.759784	4.16291	121	121	2.53293e-14	[0.66 0.83]

(b) ICC between rater 1 and rater 3

Type	Description	ICC	F	df1	df2	pval	CI95%
ICC1	Single raters absolute	0.255851	1.68763	441	442	2.31912e-08	[0.17 0.34]
ICC2	Single random raters	0.307763	2.11363	441	441	4.08172e-15	[0.16 0.44]
ICC3	Single fixed raters	0.357664	2.11363	441	441	4.08172e-15	[0.27 0.44]
ICC1k	Average raters absolute	0.407455	1.68763	441	442	2.31912e-08	[0.29 0.51]
ICC2k	Average random raters	0.470671	2.11363	441	441	4.08172e-15	[0.27 0.61]
ICC3k	Average fixed raters	0.526881	2.11363	441	441	4.08172e-15	[0.43 0.61]

(c) ICC between rater 1 and rater 4

Type	Description	ICC	F	df1	df2	pval	CI95%
ICC1	Single raters absolute	-0.436916	0.39187	635	636	1	[-0.5 -0.37]
ICC2	Single random raters	0.104523	2.47356	635	635	1.62443e-29	[-0.06 0.32]
ICC3	Single fixed raters	0.424221	2.47356	635	635	1.62443e-29	[0.36 0.49]
ICC1k	Average raters absolute	-1.55187	0.39187	635	636	1	[-1.98 -1.18]
ICC2k	Average random raters	0.189264	2.47356	635	635	1.62443e-29	[-0.13 0.49]
ICC3k	Average fixed raters	0.595724	2.47356	635	635	1.62443e-29	[0.53 0.65]

(d) ICC between rater 2 and rater 3

Type	Description	ICC	F	df1	df2	pval	CI95%
ICC1	Single raters absolute	0.010671	1.02157	156	157	0.446916	[-0.15 0.17]
ICC2	Single random raters	0.240776	2.63766	156	156	1.4707e-09	[-0.08 0.51]
ICC3	Single fixed raters	0.450196	2.63766	156	156	1.4707e-09	[0.32 0.57]
ICC1k	Average raters absolute	0.0211167	1.02157	156	157	0.446916	[-0.34 0.29]
ICC2k	Average random raters	0.388106	2.63766	156	156	1.4707e-09	[-0.18 0.67]
ICC3k	Average fixed raters	0.620876	2.63766	156	156	1.4707e-09	[0.48 0.72]

(e) ICC between rater 2 and rater 4

Type	Description	ICC	F	df1	df2	pval	CI95%
ICC1	Single raters absolute	0.257354	1.69307	165	166	0.000384635	[0.11 0.39]
ICC2	Single random raters	0.356098	2.88422	165	165	1.53029e-11	[0.03 0.58]
ICC3	Single fixed raters	0.485096	2.88422	165	165	1.53029e-11	[0.36 0.59]
ICC1k	Average raters absolute	0.409358	1.69307	165	166	0.000384635	[0.2 0.56]
ICC2k	Average random raters	0.52518	2.88422	165	165	1.53029e-11	[0.05 0.74]
ICC3k	Average fixed raters	0.653286	2.88422	165	165	1.53029e-11	[0.53 0.74]

(f) ICC between rater 3 and rater 4

Figure 30: The ICC scores between all raters.