An Item Response Theory Framework for Persuasion

Anastassia Kornilova **Daniel Argyle** **Vlad Eidelman**
FiscalNote Research
anastassia, daniel, vlad@fiscalnote.com

Abstract

In this paper, we apply Item Response Theory, popular in education and political science research, to the analysis of argument persuasiveness in language. We empirically evaluate the model's performance on three datasets, including a novel dataset in the area of political advocacy. We show the advantages of separating these components under several style and content representations, including evaluating the ability of the speaker embeddings generated by the model to parallel real-world observations about persuadability.

1 Introduction

Persuasion is the art of instilling in someone a given belief or desire to take a given action. The action can be expressing agreement with the speaker in a debate (Durmus and Cardie, 2019), making a donation to a crowdfunding campaign (Yang et al., 2019) or non-profit (Wang et al., 2019), or a Supreme Court ruling (Danescu-Niculescu-Mizil et al., 2012). Social psychology frameworks for understanding persuasion, such as the Elaboration Likelihood Model (ELM), argue that attributes of successful persuasion fall into three groups: (1) message, the text of the argument; (2) audience; and (3) speaker, the source of the argument. (Petty and Cacioppo, 1986; Lukin et al., 2017; Cialdini, 2009).

Although much attention has been given to studying the text, text in isolation fails to capture how the audiences' prior beliefs and predispositions can affect their response to the same argument. Several recent studies have considered all three factors within the context of specific datasets by creating features to represent the audience as a whole or by building separate models for different types of audiences (Lukin et al., 2017; Tan et al., 2016; Durmus and Cardie, 2019; El Baff et al., 2020). In this paper, we present a broad framework that can represent individual audience members in one model across a diverse set of persuasion tasks.

Since implementing the ELM framework requires separate data about the speaker, audience, and argument, it is difficult to validate empirically. Often, we only have access to the observed outcome (e.g. did the person donate money). Both the persuadability of the audience and the persuasiveness of the argument are unobserved. Motivated by this, we explicitly model a persuasive scenario as a function of latent variables describing the persuadability of the audience and the persuasiveness of the text.

Our approach is based on Item Response Theory (IRT), a framework for modeling the interaction between latent traits and observable outcomes. While these types of models are well known in the context of education (Fischer, 1973; Lord, 1980; McCarthy et al., 2021) and politics (Clinton et al., 2004), to our knowledge this is the first application of an IRT model to study persuasion. Using this framework, we model the interaction between the grouped *argument* and *speaker*, and the *audience*, explicitly. The argument and speaker are grouped together because in practice it is hard to separate their effects, especially in the written tasks covered in this study.

We explore two variations on the IRT framework and apply it to three different persuasion tasks. In addition to two previously studied tasks, we introduce a novel setting related to political advocacy group campaigns, where a recipient is asked by an organization to take a specific action.

We evaluate these models with different parameterizations, including style and content features, showing that they are both effective for predicting persuasion, and have the ability to uncover latent characteristics of the audience that were modeled explicitly in previous works.

Our contributions are as follows: 1) we formalize the use of IRT model formulations for persuasion and show the advantages of them over existing approaches, 2) we introduce a new dataset of

Findings of the Association for Computational Linguistics: NAACL 2022, pages 77 - 86
July 10-15, 2022 ©2022 Association for Computational Linguistics

political advocacy emails, 3) we apply the formulations with style and content features on three persuasion tasks, and 4) we show that the separate latent audience component is interpretable and consistent with external information. All code associated with the paper is available at `https://github.com/akornilo/IRT_Persuasion`.

2 Item Response Theory

Item Response Theory (IRT) represents a set of models that explain an observed outcome based on latent traits. These models are frequently used when an outcome is easily observed, but the factors predicting that model are unobservable. For example, in education an outcome could be a student's answer to an exam question, and the latent predictive traits are a students knowledge and the difficulty of the question; in politics an outcome could be a vote on a bill and the unobservable traits are the legislator's and bill's ideology. Crucially, an IRT model provides both a prediction of the outcome, and an interpretable measurement of the latent variables.

In applying IRT to persuasiveness, we can view the audience as having a response to the item, where the item is an argument composed of the speaker and message pair.

2.1 Rasch Testing Model

We build on two specific IRT parameterizations. The first, the **Rasch** model (Rasch, 1960) is commonly used in education research to model the difficulty of standardized test questions (Fischer, 1973; Lord, 1980). In it the probability that an individual i answers test question j is given by:

$$p(y_{ij} = 1 \mid \alpha, \beta) = \sigma(\alpha_i - \beta_j) \qquad (1)$$

where α_i represents a respondent (e.g. a student's ability) and β_j represents the item (e.g. the difficulty of a test question). Intuitively, if the ability is greater than the question difficulty, then the student will answer the question correctly. Given a series of exam sessions one can estimate values of α and β for all of the students and questions in the dataset. This can be done using a variety of optimization strategies, such as Expectation Maximization or Bayesian techniques (Bock and Aitkin, 1981; Natesan et al., 2016).

However, one limitation of this approach is that it cannot be used to perform inference on new test questions because all parameters are estimated simultaneously. To solve this problem, Fischer (1973) proposed the linear logistic test model that parameterizes the difficulty, β, as a weighted linear combination of test features. In this formulation, the student (α) remains a latent variable, but the β of an unseen question can be predicted using attributes of the question itself.

Following Fischer (1973), the parameterization used to predict the item parameters is a weighted linear sum of features:

$$\beta_j = \sum_{k=1}^{K} w_k \times \psi_{jk} \qquad (2)$$

where ψ_k is an input feature representing the item, and w_k is the associated weight.

In order to apply this model to persuasion, we propose considering argumentation as follows: First, arguments can vary in quality, similar to test questions having different difficulty levels. Second, we can only measure the quality of an argument based on how the audience reacted; similar to how a students ability is measured via their performance. Third, it is possible that a good argument is matched with an audience reticent to persuasion, similar to a good student receiving a particularly hard question. Note that this requires an audience member observe multiple arguments, and that each argument be heard by multiple audience members. Inspired by the linear logistic model, we model the latent argument parameter as a function of attributes of the argument itself, thus allowing us to include attributes of the speaker and text in the model directly.

2.2 Two Parameter IRT

While the simplicity of the Rasch model is powerful, a two parameter generalization of an IRT model (a two parameter logistic - **2PL**) provides additional benefits for our application (Birnbaum, 1968). In the simplest version, a two parameter model (so called because the item is modeled with two parameters) is as follows:

$$p(y_{ij} = 1 \mid \boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\beta}) = \sigma\left(\alpha_i \cdot \phi_j - \beta_j\right) \qquad (3)$$

where as before, α_i represents the respondent (students ability), and β_j is the item's difficulty,[1] but

[1] Analogous to the Rasch model, this tells us the overall difficulty level of the question

now ϕ_j represents the item's discrimination.[2] We similarly generalize this model by estimating the two item parameters, β_j and ϕ_j, as linear functions of features as in Equation 2.

This framework has commonly been be used to explain legislator voting behavior (Clinton et al., 2004), a useful analogy as many of the persuasion contexts we consider have political undertones. In this case, the response y_{ij} is a vote by respondent i (a legislator) on item j (a bill). Clinton et al. (2004) show that the parameter α_i can then be interpreted as the respondent's ideology (e.g negative values are more liberal, positive values are more conservative); ϕ_j is referred to the bills polarity (i.e. discrimination);[3] β_j represents the bill's popularity (i.e. difficulty).[4] Persuasion is a generalization of this framework because popularity can correspond to properties of arguments that are appealing overall, while polarity represents techniques or topics that appeal only to a subset of the audience.

2.3 Audience Analysis

Once a Rasch or a 2PL model is fit, the learned α can be interpreted as a one-dimensional respondent embedding. In the legislator voting context these values can be interpreted as ideologies: legislators with very negative or very positive embeddings reflect very liberal and conservative stances, respectively, while those with small-value embeddings map to moderate legislators. While interpretation of these values will depend on the task, in general, similar embeddings will map to similar audience members and can be grouped together for further analysis.

3 Related Works

Audience Effects The properties of the audience in relation to argument persuasiveness have previously been examined in several predictive studies. Lukin et al. (2017) show that audiences with a more "open" personality respond better to emotional arguments, while El Baff et al. (2020) show that liberals are more affected by the style of a new editorial than conservatives. Wang et al. (2019) also find that people with different personality types respond differently to emotional vs. logical appeals. Tan et al. (2016) show how "malleable" different Reddit users are to new perspectives. Durmus and Cardie (2018, 2019) show that prior beliefs play a strong role in how persuadable someone is. Cano-Basave and He (2016) study persuasiveness of style in political speeches. In contrast to these studies, our method is designed to work when we have limited or no information about the audience of an argument.

Item Response Theory As described in the previous section, IRT models have primarily been applied in politics to measure the ideology of politicians (Clinton et al., 2004; Poole and Rosenthal, 1985). While most IRT implementations here rely only on the responses as data, more recent work augment the models to take advantage of the text through a simultaneously estimated topic model (Gerrish and Blei, 2012; Vafa et al., 2020; Lauderdale and Clark, 2014).

The efficacy of IRT has been applied on large-scale datasets to verify the validity of standardized tests both in the U.S. and internationally (AERA et al., 2014; Rutkowski et al., 2014). Recent advances have focused on polytomous test questions and creating new questions (the 'cold-start' problem: Settles et al., 2020; McCarthy et al., 2021). In this paper, we focus on the simplest form, but this area of research points to many possible extensions.

Argument Quality Argument mining has been studied in various domains (Palau and Moens, 2009). Most relevant here, several studies have attempted to study argument quality through pairwise ranking as the outcome (Habernal and Gurevych, 2016; Gleize et al., 2019; Toledo et al., 2019).

Framing Theory In the study of framing effects, the expectancy value model (Chong and Druckman, 2007) represents an attitude as $\sum_i v_i \times w_i$, where v_i is the favorability of the object of evaluation (e.g. a candidate), on dimension i (e.g. foreign affairs or personality), and w_i is the salience weight ($\sum_i w_i = 1$). Our parameterization of β_j and ϕ_j can be seen in this paradigm as identifying frames in communication, with each feature of the style and content as a dimension, and learning the framing effect of each.

[2] Discrimination is how well the question is able to tell which students perform better, a high value indicates clearly separates high scoring students from low scoring, a negative value would indicate that low performing students are more likely to get the question right than high performing.

[3] Large negative or positive values indicate that a bill is strongly ideological, a value close to zero means the vote isn't strongly driven by ideology.

[4] Large values indicate a bill that is "difficult" to vote for and is less likely regardless of ideology.

4 Datasets

In order to apply the IRT framework, an audience member must respond to multiple arguments (and arguments must be observed by multiple audience members). Too few responses implies that an audience member's latent value will be driven entirely by the one or two arguments. While not many existing argument mining datasets meet this criteria, we are able to study three diverse settings. Additionally, our advocacy task is akin to many real-world settings where users on one-platform are asked to complete an arbitrary task (e.g. a retail mailing list getting users to click on a promotion).

4.1 NYTimes Editorials

The NYTimes Editorial corpus[5] consists of 975 editorials from the New York Times news portal (El Baff et al., 2018). Each publication was reviewed by 3 conservatives and 3 liberals from a pool of 12 conservative and 12 liberal reviewers.

Each reviewer rated the editorials as either 'challenging', 'reinforcing' or 'no effect'. These labels must be approached with care as reinforcing could imply 'reinforced view against the article's stance'. El Baff et al. (2020) study this corpus in a ternary setting by aggregating the liberal and conservative votes and building separate models for each side. For our study, we construct a binary task for predicting 'whether this article had an effect'. While this framing elides whether the speaker succeeded according to her intent, it does relay whether the argument was persuasive.

4.2 Debates (DDO) Corpus

DDO is a corpus of 78k debates scraped from `debate.org`.[6] Each debate has two speakers and an audience votes on a winner.[7] In addition, each audience member can fill out their profile with their political and religious ideology, and stance on various political issues (e.g. Abortion or the Border Wall). Originally, it was used to study how prior beliefs and similarities between the audience and the speaker affected debate outcomes (Durmus and Cardie, 2018, 2019).

To preprocess the data, we removed all debates that have fewer than three rounds, end in a forfeit or a tie, have fewer than 100 words per side, or have fewer than 5 points awarded total. In addition, we excluded debates not on the following issues: Politics, Religion, Society, Philosophy, Education and Economics. Since we are interested in modeling individual audience members, we identify audience members who have responded on at least 10 debates, then remove debates where none of those members responded. The final dataset contains approximately 60k datapoints; 6320 debates and 1131 responders.

Each debate has one side with a pro argument and one side with a con argument, resulting in the wining side being "assigned more points". The prediction task consists of whether a responder gave more points to a given debate side. Since our models only consider one argument at a time, we treat each side of the debate as a separate item, concatenating texts from all rounds from that speaker.[8]

4.3 Advocacy Campaign Corpus

Grassroots advocacy is the process wherein organizations (e.g. companies, non-profits, coalitions) encourage individual citizens to influence their government. In the United States, such lobbying often takes the form of advocacy email campaigns, sent by an organization to specific audiences, asking them to take an action, such as contacting their legislators to vote yes or no on a particular bill.

We construct a dataset containing the text and metadata of these emails, from a popular advocacy software platform, paired with whether recipients took the requested action.[9] Organizations will send different messages to the same audience over time, allowing us to identify which emails (items) elicited a response from specific recipients. Thus, it is possible to distinguish messages that did not generate interest overall (popularity) from messages that did not resonate with specific groups of recipients (polarity).

The dataset contains 63,795 individual recipients of 7,067 email campaigns from 328 different organization, resulting in approximately 2 million individual data points. Each recipient has data for

[5] `https://webis.de/data/webis-editorial-quality-18.html`

[6] `https://www.cs.cornell.edu/~esindurmus/ddo.html`

[7] While the audience can assign points to various aspects of the debate, this study will only consider the cumulative sum of the points.

[8] We are interested in how a single unit of argument affects the audience, and leave extension of this to account for both simultaneously to future work.

[9] Due to privacy concerns, this dataset will not be released, but platform users agreed to terms of services providing for internal analysis.

15 to 100 emails and had an action rate between of 5% - 95%.[10] Each email included in the dataset had at least 6 responses.

The data is not balanced with respect to organizations; while the largest organizations sent over 200 emails, the median is 6. One possibility of this imbalance is overfitting a feature that is only pertinent to one, particularly prevalent organization. To mitigate such effects, we include an indicator variable to specify the organization.[11]

5 Model Features

Argument analysis is often separated into *style* and *content* features (Cano-Basave and He, 2016; Longpre et al., 2019; El Baff et al., 2020), with additional categories included for argument quality and task specific properties. Since we group the speaker and the argument text together, we combine features representing both as inputs to ϕ and β.

Lexicon Style Features Style features represent higher-level properties of words and rhetorical structures. We chose the following sets of such features from lexicons that were commonly used in previous argumentation literature:

LIWC lexicon of 93 metrics ranging from parts-of-speech to thinking styles to emotions (Pennebaker et al., 2015);[12] Valence, Arousal, Dominance (Warriner et al., 2013); Concreteness (Brysbaert et al., 2014). (These features were shown to be useful for argument quality analysis by Tan et al. (2016).) Argument features developed by Somasundaran et al. (2007), including *necessity, emphasizing, desire, contrasting* and *rhetorical question*; NRC Lexicon: Word-level level associations for emotions like anger, disgust and fear (Mohammad and Turney, 2013); Sentiment and Subjectivity: as implemented in the TextBlob Python Library.[13]

Argument Text We use TF-IDF unigrams to represent the text directly (tuned with respect each task). While we initially explored using deep, contextual text representations, they did not show benefit, and the motivation for this paper is to understand the benefits of the IRT framework, rather than optimize performance based on the argument alone.

Debate-Only Speaker Features In the debate platform, users can optionally specify a stance - for, against, undecided or no stance - on 48 issues such as Abortion, Death Penalty or Gay Marriage. These can be viewed as a proxy for the content as users often present arguments that align with their views.

Advocacy-Only Org Indicator An indicator to account for the large variation in action rate between organizations. Additional indicators are used to represent the industry and organization size.

Advocacy-Only Appeals Using data from Wang et al. (2019), we built a multi-class classifier to recognize ‘emotional’, ‘logical’ and ‘credibility’ appeals. The classifier was applied at a sentence level to the emails, and features were created for the average and the sum of the scores across the sentences.

Advocacy-Only Misc Features : The day of the week and time of day have a strong effect on email click rate.[14] We include indicator features for the day of the week and the hour of day. We include an *urgency* indicator feature, based on a custom list of words indicative of high urgency and timeliness (e.g. “soon”, “now”, “hurry”).

IBM Quality Gretz et al. (2019) released a dataset of 30k sentence-level arguments with 0-1 quality ratings. Unlike our tasks where quality is a latent property, these sentences were assessed for quality directly. We re-implemented the BERT-FT model from this paper, using the MACE-P score. Since these scores were trained on short texts, we apply them to individual sentences in the input text, then use the min, max, average, range, 25th, 50th, and 75th percentiles of these scores. As far as we know, this is the first study to transfer the quality model to longer texts. These features will be grouped with Style for the analysis.

[10] Those with a lower or higher action rate are unlikely to be illustrative of persuasion characteristics.

[11] Alternatively, we could construct separate models for each organization, but refrain from doing so for three reason. First, about a quarter of recipients are ‘multi-org’ - they receive emails from multiple sources, thus, we would like to model their behavior across all of them. Second, as many of the organizations are not well represented, they benefit from patterns that appear across different organizations. Finally, maintaining a separate model for every recipient and recipient is not as efficient or scalable.

[12] We purchased a copy of the software from liwc.wpengine.com to obtain these labels.

[13] https://textblob.readthedocs.io/

[14] https://sleeknote.com/blog/best-time-to-send-email

Model	Accuracy
Audience Prior	0.662
Style	0.741
Text	**0.754**
Style + Text	0.750

Table 1: Results for the Editorials Task (Rasch Model).

6 Models and Results

Since the Editorials corpus is the smallest, we use the simpler Rasch parameterization, while the 2PL model is used for the Debates and Advocacy tasks.[15] Each of the models is trained using a regularized binary cross-entropy loss:

$$L\left(\hat{y}_i, y_i\right) = -y_i \log \hat{y}_i - \left(1 - y_i\right) \log\left(1 - \hat{y}_i\right) \\ + c \cdot \|\alpha, \beta, \phi\|$$

where $\hat{y}_i$ is the output from equation 1 or 3, and y_i is the binary label, representing if the persuasion was successful. The second part of the equation represents a regularization parameter. Details on the experimental parameters can be found in Appendix A. For each task, an audience prior baseline is used. It is generated by calculating the rate at which the audience member was persuaded in the training data (e.g. did the article have an effect, how many recipients took the requested action), then drawing labels on the test data accordingly.

6.1 Editorial Results

The results on the Editorial task are shown in Table 1. The performance for all three feature sets is relatively similar, with all outperforming the audience prior.

The embeddings and weights generated by the model can be analyzed separately for further insights. First, in Figure 1 we compare the distribution of audience embeddings (α) for the liberal and conservative reviewers. According to our theory, these can be interpreted as individuals reticence to being persuaded. While a majority of reviewers are close to 0, we see two liberals with larger negative values (meaning they are particularly open to the messages) and several conservatives on the right (suggesting they are more closed off to these messages). This supports El Baff et al. (2020) observation that conservatives are generally resistant to the *New York Times style*; however, the fact that the majority of reviewers from both sides have similar embeddings, suggests that the pattern is not very strong.

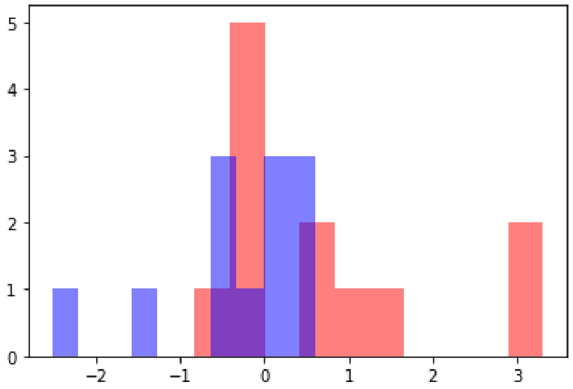

Figure 1: Reviewer Embeddings for the Editorial Rasch Model on the x-axis. Blue represents liberal reviewers, red represents conservative reviewers.

This data also contained information from each reviewers Big 5 Personality test. We measured the Pearson correlation between the reviewers embeddings and found a strong correlation with extroversion (r=-0.568, p<0.05) and openness (r=-0.344, p<0.1). These findings closely match El Baff et al. (2018)'s analysis between Big 5 Personality Ratings and the affectedness labels. The audience embedding is a latent parameter, thus, it does not explicitly represent personality or political preferences. This analysis has two implications: first, the IRT framework is successful in situations where additional data about the audience is not available; second, analyzing the embeddings lets us learn qualities of the audience post-hoc.

For style, the highly weighted features included negative sentiment markers (*nrc_negative, liwc_negative_emotions*); this aligns with El Baff et al. (2020)'s observation that ineffective editorials tend to have a neutral tone (although their study only focuses on liberal reviewers). The quality features do not show consistent behavior: the *quality_mean* feature has a large negative weight (e.g sign of a bad editorial), but the 75th and 25th percentile features have positive weights; suggesting that the quality measure does not transfer well to editorials.

6.2 Debate Results

The Debates data is approximately 10 times larger than Editorials and contains a more diverse audience. The results are shown in Table 2. Without the

[15] In addition, there is natural polarity in the Debate task that lends itself to the 2PL model, as ϕ in equation 2 is designed to model such an effect.

Model	Accuracy
Random	0.500
Style	0.561
Text	0.581
Speaker	0.611
Speaker + Style	0.626
-β (popularity) layer	0.604

Table 2: Results for Debates Task (2PL Model).

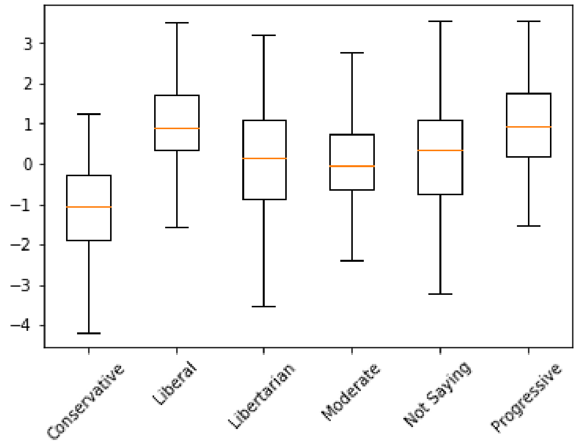

Figure 2: Distribution of one-dimensional audience embeddings on the y-axis.

popularity parameter, β the performance decreases, which confirms the theory that both polarity and popularity are necessary to adequately represent the argument and the speaker. The Speaker stance model outperforms just Text; a probable explanation is that the stances are a proxy for the actual opinions expressed in the text that a simple unigram representation can not capture.

To understand the latent audience embeddings we compare them to the self-reported political affiliations from their profiles. Figure 2 shows a clear separation between liberals and conservatives (the two largest groups). This finding supports the work of Durmus and Cardie (2019) which showed that similarity on 'Big Issue Stance' between the speaker and the audience member is a good indicator for predicting outcome. As with Editorials, the advantage of our approach is that we were able to infer audience member preferences without using their profiles.

To understand what ϕ and β tells us about persuasive theory, we will focus on the Speaker+Style model:

High Polarity: Abortion, Gay Marriage, Progressive Tax;

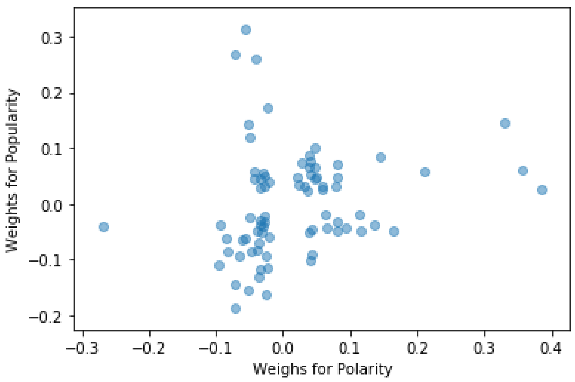

Figure 3: Contrast of weights from popularity vs polarity features.

Low Polarity: Border Fence, Gun Rights, Homeschooling;

High Popularity: quality_max, quality_range, liwc_differ;

Low Popularity: liwc_Exclam, liwc_authentic, liwc_drives

For popularity the significant factors are related to style and quality. The high 'quality_max' feature suggests that the quality model transfers better to this context than Editorials. The low popularity value for 'liwc_authentic' is interesting, as El Baff et al. (2020) also found that authenticity generally led to No Effect editorials.

For polarity, the highest weighted are the stances. 'Polarity High' corresponds to having a Pro stance on those issues, which in this case represent a Liberal view point. This corresponds with the Liberal recipient embeddings in Figure 2 having generally positive embeddings (alignment in weights results in positive final weight). The opposite is true for the Conservative issues and embeddings. This alignment reinforces the finding that prior beliefs play a strong role in outcomes (Durmus and Cardie, 2018).

Figure 3 plots the weights learned for each feature for the polarity and popularity parameters.[16]

Notably, the orthogonal pattern extends beyond the top features, features that strongly predict whether the audience responds to an argument do not usually strongly predict whether the argument is popular overall.

[16] This figure excludes features that had very small weights along both dimensions.

	Overall		Audience Average		Org Average	
	Acc.	Macro-F1	Acc.	Macro-F1	Acc.	Macro-F1
Org Prior	0.608	0.514	0.606	0.263	0.630	0.513
Audience Prior	0.710	0.415	0.716	0.318	0.714	0.472
Org Only	0.757	0.667	0.759	0.589	0.728	0.573
Org + Style	**0.781**	**0.708**	**0.761**	**0.662**	**0.771**	**0.678**
- β (popularity)	0.750	0.653	0.749	0.643	0.756	0.654
Sep Feat V1	0.725	0.619	0.726	0.571	0.700	0.520
Sep Feat V2	0.748	0.678	0.750	0.604	0.698	0.654

Table 3: Results For Advocacy Task (2PL Model).

6.3 Advocacy Results

Table 3 shows the results for the Advocacy task.[17] The overall accuracy and macro-F1 scores represent results across all data, while the Org and Audience average accuracy represent data for individual organizations and respondents. Due to the variation in action rate and sample size, the macro-F1 results are particularly important.

While the Org Only model performs well,[18] the improved performance with the additional of Style suggests that the style of an email still affects the user. The style features may have an advantage for recipients associated with a diverse set of organizations. Without β, the performance is significantly worse, again confirming the need for both parameters.

To better understand the effect of style and org features, two additional models are trained that separate between polarity and popularity. In **Sep Feat V1**, ϕ receives style features, β receives org indicators. In this setting, $(\alpha \cdot \phi)$ represents how individuals are affected by style, while β models the organizations base rate. In **Sep Feat V2** the features are reversed. V1 has the worst performance of all five 2PL models, suggesting that modeling the interaction between the recipient and organization $(\alpha \cdot \phi)$ is important. Org-Only and V2 have mixed performance on accuracy, but V2 performs better on macro-F1, suggesting that style influences the recipients' decisions to act.

Finally, we analyze the features with lowest and highest magnitudes from β in the Org+Style model. The highest weighted features include *concreteness, average-logical-appeal, word count* and *quality 75th percentile*. The lowest weighted features (unlikely to produce action) include *valence, quality mean, arousal* and *liwc-we*. Similar to the Editorials, the quality features are contradictory, suggesting the connection between sentence level and document level quality needs to be investigated further. The logical appeal feature shows they are particularly effective (the corresponding scores for emotional and credibility appeals had smaller, negative weights).

7 Conclusion and Future Work

In this paper, we validate the social psychology frameworks for persuasion using the IRT framework to explicitly model the audience and the speaker. Our approach lets us analyze how different audience members respond to the same argument, and we show that our representation implicitly learns latent audience features modeled explicitly by other models.

We empirically showed several additional insights about persuasion. In the Debates and Advocacy tasks, the Popularity parameter improved performance showing that certain stylistic elements are universally appealing. In the Debates task, the audiences' embeddings aligned with their political affiliation, showing that prior beliefs play a strong role in their argument perception. While the background information about the audiences was available for these tasks, we did not need to model it explicitly; as a result this setup allows us to make predictions for audiences who do not report their affiliation.

A potential negative side of the models is they may learn latent characteristics of the speaker or

[17] Due to computational constraints, we omitted the raw text model from this task.

[18] One likely explanation for this performance is that audience is not independent of the speaker - by virtue of receiving emails from this organization, recipients may also have similar preferences.

audience they may not be aware of or consider private. However, all datasets studied in this paper were either public and anonymous or private with audiences who consented to analysis.

This study focused on simple representations to show the viability of our method and provide for explainability. To build on this foundation in future work, we will: expand argument text representations with contextual word embeddings and stance detection models; include higher dimensional embedding for audience and item parameters (the IRT models easily generalize to this set-up). These improvements will allow us to better capture the elements of persuasion, especially in a complex case like Advocacy.

References

AERA, APA, and NCME. 2014. *Standards for Educational and Psychological Testing*.

A. L. Birnbaum. 1968. Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*.

R Darrell Bock and Murray Aitkin. 1981. Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, 46(4):443–459.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911.

Amparo Elizabeth Cano-Basave and Yulan He. 2016. A study of the impact of persuasive argumentation in political debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1405–1413, San Diego, California. Association for Computational Linguistics.

Dennis Chong and James N. Druckman. 2007. Framing theory. *Annual Review of Political Science*, 10(1):103–126.

R.B. Cialdini. 2009. *Influence: The Psychology of Persuasion*. Collins Business Essentials. HarperCollins e-books.

Joshua Clinton, Simon Jackman, and Douglas Rivers. 2004. The statistical analysis of roll call data. *American Political Science Review*, 98(2):355–370.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of WWW*, pages 699–708.

Esin Durmus and Claire Cardie. 2018. Exploring the role of prior beliefs for argument persuasion. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Esin Durmus and Claire Cardie. 2019. A corpus for modeling user and language effects in argumentation on online debating. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 602–607, Florence, Italy. Association for Computational Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2018. Challenge or empower: Revisiting argumentation quality in a news editorial corpus. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 454–464, Brussels, Belgium. Association for Computational Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2020. Analyzing the persuasive effect of style in news editorial argumentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3154–3160, Online. Association for Computational Linguistics.

Gerhard H. Fischer. 1973. The linear logistic test model as an instrument in educational research. *Acta psychologica*, 37(6):359–374.

Sean M. Gerrish and David M. Blei. 2012. The issue-adjusted ideal point model.

Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2019. Are you convinced? choosing the more convincing evidence with a siamese network. *CoRR*, abs/1907.08971.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2019. A large-scale dataset for argument quality ranking: Construction and analysis. *CoRR*, abs/1911.11408.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.

Benjamin E. Lauderdale and Tom S. Clark. 2014. Scaling politically meaningful dimensions using texts and votes. *American Journal of Political Science*, 58(3):754–771.

Liane Longpre, Esin Durmus, and Claire Cardie. 2019. Persuasion of the undecided: Language vs. the listener. In *Proceedings of the 6th Workshop on Argument Mining*, pages 167–176, Florence, Italy. Association for Computational Linguistics.

Frederic M Lord. 1980. *Applications of item response theory to practical testing problems*. Routledge.

Stephanie M. Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument strength is in the eye of the beholder: Audience effects in persuasion.

Arya D. McCarthy, Kevin P. Yancey, Geoff T. LaFlair, Jesse Egbert, Manqian Liao, and Burr Settles. 2021. Jump-starting item parameters for adaptive language tests. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 883–899, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Saif Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *CoRR*, abs/1308.6297.

Prathiba Natesan, Ratna Nandakumar, Tom Minka, and Jonathan D Rubright. 2016. Bayesian prior choice in irt estimation using mcmc and variational bayes. *Frontiers in psychology*, 7:1422.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.

Richard E Petty and John T Cacioppo. 1986. The elaboration likelihood model of persuasion. In *Communication and persuasion*, pages 1–24. Springer.

Keith T Poole and Howard Rosenthal. 1985. A spatial model for legislative roll call analysis. *American Journal of Political Science*, pages 357–384.

G. Rasch. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Studies in mathematical psychology. Danmarks Paedagogiske Institut.

Leslie Rutkowski, Matthias Von Davier, and David Rutkowski. 2014. *Handbook of international large-scale assessment. Background, technical issues, and methods of data analysis.*

Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. Machine learning–driven language assessment. *Transactions of the Association for Computational Linguistics*, 8:247–263.

Swapna Somasundaran, Josef Ruppenhofer, and Janyce Wiebe. 2007. Detecting arguing and sentiment in meetings. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 26–34.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of WWW*.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment - new datasets and methods. *CoRR*, abs/1909.01007.

Keyon Vafa, Suresh Naidu, and David M Blei. 2020. Text-based ideal points. *arXiv preprint arXiv:2005.04232*.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

Diyi Yang, Jiaao Chen, Zichao Yang, Dan Jurafsky, and Eduard Hovy. 2019. Let's make your request more persuasive: Modeling persuasive strategies via semi-supervised neural nets on crowdfunding platforms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3620–3630, Minneapolis, Minnesota. Association for Computational Linguistics.

A Model Training Details

The models described in section 6 were trained as follows. In equation (6), c is set to $1e^{-4}$ for all experiments. $L2$ loss is used for the Editorials and Advocacy corpus and text model for Debates, $L1$ is used for the remaining models in the Debates corpus. Editorial models are trained for 200 epochs; Debates for 25; Advocacy for 5. A learning rate of 0.01 is used for Editorials and Debates; 0.005 is used for Advocacy.

All results are reported over 5-fold cross-validation, with the splits performed at an argument level. All models are fit using the AdamW optimizer. The α embedding initializations are drawn from a uniform distribution of -0.5 to 0.5.