Words Matter: The Role of Readability, Tone, and Deception Cues in Online Credit Markets Internet Appendix

A1. Related Literature

Our study is most closely related to Duarte, Siegel, and Young (2012), Iyer, Khwaja, Luttmer, and Shue (2016), Dorfleitner, Priberny, Schuster, Stoiber, Weber, De Castro, Kammler (2016), and Ravina (2019). As discussed in our manuscript, the Iyer et al. (2016) study takes an innovative approach to examining whether Prosper investors use soft data. Specifically, the authors estimate the information contained in soft, non-easily quantifiable data as the residual from a regression of the interest rate on the hard credit variables and the easily quantifiable nonstandard information (e.g., the maximum rate a borrower is willing to pay). As discussed in our study, the easily quantifiable nonstandard information includes a number of textual dimensions (e.g., average sentence length). The Iyer et al. (2016) indirect estimate, however, will also capture information lenders infer from hard credit and easily quantifiable nonstandard data unless the model is perfectly specified. That is, variation in the left-hand side (interest rate) attributed to hard credit or easily quantifiable nonstandard data will end up in the residual if the model is less than perfectly specified. Because this indirect estimate (i.e., residual) helps explain default risk (their Table 5), the authors infer that peer-to-peer lenders use non-easily quantifiable soft data (i.e., borrowers' photo and writing) to more efficiently price loans.

Two of the other closely related studies focus on investors inferring information from Prosper borrowers' photos and find conflicting evidence. Duarte et al. (2012) conclude, consistent with the Iyer et al. (2016) indirect evidence, that borrowers who appear "trustworthy" are more likely to be funded, charged a lower rate, and less likely to default. In contrast, Ravina (2019) concludes, inconsistent with the Iyer et al. (2016) indirect evidence, that "beautiful" borrowers are more likely to be funded and charged lower rates, but just as likely to default. As discussed in our study, although both Duarte et al. (2012) and Ravina (2019) focus on soft information contained in images, both include a number of easily quantifiable linguistic metrics.

In contrast to Duarte et al. (2012) and Ravina (2019) who focus on soft information inferred from photos, Dorfleitner et al. (2016) focus on borrowers' writing. The authors hypothesize that three linguistic dimensions—spelling errors, text length, and indicators for four sets of social and emotion words—will predict both funding and default likelihood. Specifically, the authors propose that fewer spelling errors and medium text length (positively related "up to a certain amount of words" before becoming negatively related) will be associated with higher funding likelihood and lower default probability. In contrast, the authors posit that investors will respond irrationally to social and emotional words (e.g., investors will want to help someone going through a divorce even if they offer a poor risk-return tradeoff) and therefore these indicators will be associated with both a higher funding likelihood and a greater default likelihood.

The authors use indicator variables for the presence of any one of seven "positive emotion" keywords (thank you, rejoice, dream, urgent, healthy, desire, and trust), any one of five "negative emotion" keyword (funeral, lament, sick, difficult, and deceased), any one of ten "family" keywords (wife, husband, upbringing, family, marriage, wedding, child, children, married, engagement), and any one of four "separation" keywords (divorced, two German words for divorce, and separation).[1]

Focusing on the results from their full model (i.e., including all the metrics simultaneously), the authors find no evidence to support any of their hypotheses that spelling errors, text length, or indicators for the presence of any of the four social and emotional word sets are meaningfully associated with default likelihood. That is, inconsistent with the indirect evidence in Iyer et al. (2016), Dorfleitner et al. (2016) find no evidence investors use linguistic dimensions to extract value-relevant information regarding the loan.

With respect to funding likelihood, the Dorfleitner et al. (2016) empirical results are both limited and sample dependent. For example, none of the coefficients associated with the indicators for separation or family keywords differ meaningfully from zero in the expected direction (the

[1] Translations from Google Translate.

coefficient associated with family keywords is statistically significant in one of the two platforms, but has the wrong sign). Assuming the Dorfleitner et al. (2016) indicators for positive and negative keywords capture, to some extent, tone, the authors find no evidence tone is related to default likelihood and conflicting evidence that tone is related funding likelihood, i.e., the relation between funding likelihood and tone is positive in one sample and negative in the other sample.[2]

Last, and perhaps most important, comparison across the German and U.S. peer-to-peer markets is extremely limited as a result of differences in the platforms and data. For instance, because credit scores are optional in one of platforms Dorflietner et al. (2016) examine, more than 75% of the loans on that platform have no credit data, i.e., the data do not allow the authors to control for credit groups.[3]

Related studies in Accounting and Marketing

Michels (2012) uses Prosper data to test if lenders interpret voluntary unverifiable disclosures as credible signals. Michels notes that previous work addressing this question focuses on voluntary disclosures in audited financial reports (see Healy and Palepu (2001) for a review this literature) which faces a number of issues (e.g., determining whether firm performance impacts disclosures or whether disclosures impact firm performance) that the Prosper data can overcome. Specifically, Michels selects a stratified sample of 500 funded loans and 500 unfunded loans during our sample period.[4] The author then manually assigns each listing one point for each of nine potential "disclosures": (1) the purpose

[2] Specifically, the authors find that the presence of one of the seven positive words (arguably a proxy for positive tone) is positively related to funding likelihood in their Auxmoney sample, but the presence of one of the five negative words (arguably a proxy for negative tone) is positively related to funding likelihood in Smava sample. Moreover, there is no evidence that positive word presence is related to funding likelihood in the Smava sample or that negative word presence is related to funding likelihood in the Auxmoney sample.

[3] The authors use credit score category indicators. However, as shown in their Table 2, more than 75% of the Auxmoney sample is missing credit score category data (i.e., all the credit score category indicators are zero).

[4] Michels (2012) reports his sample period ends on October 31, 2008. Our sample period ends October 16, 2008 when Prosper entered its "quiet period" (see https://www.lendacademy.com/a-look-back-at-the-lending-club-and-prosper-quiet-periods/).

of the loan (e.g., using the money to pay off credit cards), (2) income (the listing reports monthly or annual income), (3) income source (beyond the mandatory job title in the listing), (4) education (borrower discloses education level), (5) amount of debt (discloses outstanding balance of other debt), (6) interest rate on debt (reports the interest rate on at least one other debt), (7) explanation of poor credit (e.g., hospitalization), (8) monthly expenses (the listing reports a budget), and (9) a photograph. Thus, each of these 1,000 listings has a disclosure score of 0 to 9. Michels (2012) focuses on three hypotheses—more disclosures are associated with lower interest rates; more disclosures are associated with more bids; and these relations are stronger for lower credit grade loans. His empirical results support all three hypotheses.

As a "natural extension" of his primary tests, Michels (2012) examines the relation between default, his control variables, and the number of disclosures and finds a negative association between the number of disclosures and default. In contrast to the Iyer, Khwaja, Luttmer, and Shue (2016) results, however, Michels (2012) finds no evidence that lenders charge riskier loans higher rates, i.e., no evidence Prosper lenders use the soft data (disclosures) to set interest rates.

In the marketing literature, Herzenstein, Sonenshein, and Dholakia (2011) use the Prosper data to examine the role of (six possible) "identities" disclosed in the Prosper text. Specifically, the authors select a total of 1,493 Prosper listings from June 2006 (513 listings) and June 2007 (980 listings). Based on the authors' reading of these listings, they create six "identifies" (examples are from the authors' Table 1): (1) trustworthy (e.g., "I am responsible for paying my bills and lending me funds would be a good investment."), (2) successful (e.g., "I have [had] a very solid and successful career with an Aviation company for the last 13 years"), (3) hardworking (e.g., "I work two jobs. I work too much really. I work 26 days a month with both jobs."), (4) economic hardship (e.g., "Unfortunately, a messy divorce and an irresponsible ex have left me with awful credit"), (5) moral (e.g., "On paper I appear to be an extremely poor financial risk. In reality, I am an honest, decent person."), and (6) religious ("One night, the Lord awaken me and my spouse…our business has been an enormous success with G-d on our side"). The authors examine the relation between the number of identifies and three outcome variables. First, they denote the ratio of total dollar value of bids to loan request amount (ranges from 0% to 905%) as loan funding. Second, they measure the percentage difference between the borrower's stated maximum rate and the final market-clearing rate.[5] Third, they measure whether the loan defaults in the initial two years of the three-year loan term. Consistent with their hypothesis, the authors find that a greater number of identities is associated with greater loan funding and a larger "reduction" in the interest rate. Further consistent with their hypothesis, but inconsistent with the Iyer et al. (2016) indirect evidence, Herzenstein, Sonenshein, and Dholakia (2011) find that investors do not infer value-relevant information from these identities—rather, these identities mislead investors such that a greater number of identities is associated with higher default risk.

Netzer, Lemaire, and Herzenstein (2019) take a data-mining approach to find the most common individual words (out of the approximately 171,000 words in the English language), 1,052 bi-grams (two adjacent words), and 64 word categories (e.g., swear words, filler words, perception words, etc.) that are associated with defaulted versus non-defaulted loans. That is, the authors do not hypothesize any relations—they instead search for words that are more common in defaulted loans than non-defaulted loans. The authors include several additional linguistic measures: (1) the number of characters in the loan request title, (2) the percent of words with more than five letters, (3) the "Simple Measure of Gobbledygook", and (4) the number of spelling errors (unscaled by number of words). Inconsistent with the hypothesis that investors extract information from these measures, however, the authors find little evidence these variables are meaningfully related to default (only "words with more than six letters" is meaningfully related to default). The authors' data-mining

[5] Recall that during this period Prosper acted as a reverse Dutch auction where rates "started" at the highest rate a borrower was willing to pay and lenders "bid down" the interest rate.

technique identifies words that are more common in defaulted versus non-defaulted loans. For instance, defaulted loans are more likely to contain the word "God" while non-defaulted loans are more likely to contain the word "lend."

In addition to focus and method, our study differs from the related work in accounting and marketing in the questions examined. For instance, although it is possible there could be some overlap between certain "identities" or "disclosure" categories and our three linguistic metrics, our items are clearly unique from these variables. For instance, any given identity could have high or low readability, high or low deception cues, and positive or negative tone. The hard work identity, for example, could have a positive tone (e.g., "My hard work will ensure you get paid") or a negative tone (e.g., "Despite my hard work, life has been unfair"). Similarly, a direct comparison between the marketing and accounting literature and most of the work in finance on peer-to-peer lending markets is limited due to the uniqueness of the measures and control variables. For example, instead of focusing on the market-clearing rate (as in Duarte et al. (2012), Ravina (2019), Iyer et al. (2016), Liskovich and Shaton (2017)), Herzenstein et al. (2011) measure the percentage difference between the borrower's maximum rate and the market-clearing rate. Thus, for example, if two equal credit risk borrowers had a market-clearing rate of 15%, but the first borrower had a maximum rate of 20% while the second borrower had a maximum rate of 15%, their "interest rate measure" would be 25% for the first borrower ((0.20-0.15)/0.20) and 0% for the second borrower ((0.15-0.15)/0.15). In addition, although Herzenstein et al. (2011) include controls for credit grades, they do not include any of the other control variables we list in Table 1 Panels B (verified hard credit information), C (unverified credit information), or D (auction characteristics). Similarly, Netzer, Lemaire, and Herzenstein (2019) do not include most these controls in their empirical analysis. We further differ these authors in that our study develops *measures and hypotheses* based on the literature in computational linguistics, finance, MIS, and accounting. In contrast, Netzer et al. (2019) take an atheoretical data-mining approach to identify words, bi-grams, and word groups that are more common in defaulted loans relative to non-defaulted loans.

A2. Construction Details for Readability

As noted in the paper, our readability metric consists of three dimensions—spelling errors, grammatical errors, and lexical complexity. Our spelling error corpora are from two sources: the Peter Norvig spelling errors list (Jurafsky and Martin (2000)) and Birkbeck spelling error corpus (Mitto (1987)) gathered from Oxford Text Archive. The spelling error variable is defined as the ratio of the number of spelling errors to the number of words in a loan description. We randomly selected 300 loan descriptions and had one research assistant compare manually calculated scores to the machine-calculated scores. The results indicate a precision of 91% (the number of properly classified spelling errors over the number of computer-identified (correctly and incorrectly) spelling errors) and recall of 89% (the number of correctly identified spelling errors to the number of actual spelling errors).

We use a rule-based grammar check tool that employs a set of manually developed rules to match against a part-of-speech (POS) tagged text. For example, the rule of "I + Verb (3rd person, singular form)" presents an incorrect verb form usage as in the phrase "I has a cat." Specifically, we use the open source grammar checker LanguageTool.[6] It implements a set of error-catching rules that are a combination of syntactic and real-world errors found in most corpora. This tool is widely tested and used in both academia and industry (e.g., Ehsan and Faili (2013), Miłkowski (2010)). To further verify its accuracy in our context, we randomly select 500 loan descriptions and asked three experienced MTurks, who reside in the U.S. and hold a U.S. higher education degree, to examine the output. The results of the manual check suggest the LanguageTool classification achieves an accuracy of 81.3%.

The Gunning (1969) FOG index estimates the lexical complexity of texts and is perhaps the

[6] Naber (2003) provides additional detail. Also see the project home page at http://www.languagetool.org.

most commonly used financial report readability metric (e.g., Li (2008), Miller (2010), Lehavy, Li, and Merkely (2011), Dougal, Engelberg, Garcia, and Parsons (2012), Lawrence (2013), Franco, Hope, Vyas, and Zhou (2015), Asay, Elliott, and Rennekamp (2017), Hwang and Kim (2017), Lo, Ramos, and Rogo (2017)). We use the DuBay (2004) FOG formula:

$$\textit{FOG Score}=0.4 \times (\textit{Average sentence length} + 100 \times \textit{Average hard words}),$$

where average sentence length is the number of words in the description divided by the number of sentences in the description and average hard words is the proportion of words containing at least three syllables.[7] We use Natural Language Toolkit (see www.nltk.org for details) to compute the number of syllables for each word.

A3. Construction Details for Tone

We implemented a machine learning, rather than lexicon, approach to quantify positive tone. The machine learning approach requires a pre-coded training dataset (derived from manual coding) that consists of texts and their labels. By contrast, the lexicon-based approach can be faster, although an appropriate context dictionary is critical (Loughran and McDonald (2011)). Evidence (e.g., Huang, Zang, and Zheng (2014)) suggests, however, that the machine learning approach to identifying tone greatly outperform the lexicon approach.

To quantify positive tone, we begin by preparing a manually coded sample of texts from our dataset of loan descriptions. Specifically, we use stratified sampling and extract a 1% random sample of all loan request descriptions from each credit grade over the period when Prosper had borrower loan description data (February 2006 to January 2013). We partitioned each description into sentences to form our coding data set. To ensure accuracy and consistency, two research assistants coded all

[7] As noted in our manuscript, several studies (e.g., Loughran and McDonald (2014), Bonsall, Leone, Miller, and Rennekamp (2017)) point out the limitations of the FOG index in professional financial writing (e.g., annual reports). The index, however, is well-suited for examining the non-professional writing of Prosper loan applicants.

sentences in this sample dataset. Each research assistant classified each sentence as negative, neutral, or positive. The agreement rate between the two research assistants was 90% for 3,379 coded listings.

We next randomly divide the coded texts into training (70% of the total coded texts) and testing (30% of the total coded texts) sets. We considered unigram, bigram, trigrams, POS (part of speech tagging), and adjectives as potential features to identify tone. Unigrams are single content words (e.g., duck, table, thought) but exclude stop works such as "it," "a," and "the." Bigrams are adjacent word pairs. For instance, the previous sentence has bigrams of "Bigrams are," "are adjacent," "adjacent word," and "word pairs." Trigrams are three adjacent words in a sentence. POS tagging assigns the part of speech (e.g., noun, verb, adjective, adverb, etc.) to each word in a sentence. Machine learning is well suited to this task because many words must be evaluated in context (e.g., "flies" could be a noun or a verb).

We use the SVMlight multiclass machine learning package (Crammer and Singer, 2002) to train and test our data sets with all parameters set to default values (Joachims, Finley, and Yu, 2009).[8] The classifier incorporating Unigram and POS tags performed the best, achieving precision of 84% (the number of properly classified (positive, neutral or negative) tone sentences over the number of identified (correctly and incorrectly) sentences and recall of 89% (the ratio of the number of correctly identified tone sentences to the number of actual sentences). The machine learning approach generates a sentiment score (higher for more positive) for each sentence in a description. We use the average sentiment score for all sentences in each description (rescaled to zero mean, unit variance across all descriptions).

[8] See svmlight.joachims.org for additional detail.

A4. Construction Details for Deception Cues

As noted in our manuscript, the deception cue metric is computed from the fraction of: (1) exclusion words, (3) motion words, (3) first-person pronouns, (4) third-person pronouns, and (5) negative emotion words. We use the Linguistic Inquiry and Word Count (LIWC) dictionary for all definitions (see www.liwc.net). Each variable is standardized (re-scaled to zero mean, unit variance) and the raw deception cue variable is computed as standardized motion words + standardized third-person pronouns + standardized negative emotion words - standardized first-person pronouns - standardized exclusion words. We then standardize the raw deception cue variable to generate our deception cue measure.

A5. Linguistic Metrics Correlations

One potential concern is that the linguistic dimensions we examine—readability, tone, and deception cues—may be highly correlated. As shown in Table A.1, the correlation between the three metrics is quite modest—ranging from 0.04 (for readability and deception cues) to -0.13 (for deception cues and tone).

[Insert Table A.1 about here]

A6. Sample Limited to Open for Duration Listings

As noted in the manuscript, 24% of the listings are “closed when funded” (i.e., Table 2, Panel D shows that 76% of listings are “open for duration”). Because the closed when fund loans have a maximum ratio of dollars bid/dollars requested of one, while the ratio is not bound for open for duration loans, we re-examine the relation between funding outcomes (funded, number of bids, ratio of dollar value of bids to amount requested) and the linguistic dimensions when limiting the sample to open for duration loans. Table A.2 is directly analogous to Table 3 except the sample is limited to open for duration listings (and therefore excludes the open for duration indicator). The first column reports marginal effects (and associated standard errors) from a probit regression of the funding indicator on three linguistic measures and the full set of controls. The last three columns report marginal effects (and associated standard errors) from Tobit regressions of the number of bids, the ratio of total dollar amount bid to dollar amount requested, and interest rate, respectively, on the three linguistic measures and the full set of controls. The results in Table A.2 are fully consistent with those reported in Table 3 as the marginal effects are of similar magnitude and statistical significance.

[Insert Table A.2 about here]

A7. Linguistic Measures and Funding Success: Total Dollars Bid

Our primary analysis includes total dollars bid scaled by the dollar amount requested. As a robustness check, we repeat the analysis but replace the dependent variable with (unscaled) total dollar amount bid. Specifically, the first column of Table A.3 reports marginal effects from a Tobit regression of the loan's total dollar amount bid on the three linguistic measures and the full set of controls. The second column reports results from the same analysis when limited to the sample of open for duration listings.

[Insert Table A.3 about here]

The results in Table A.3 are fully consistent with the analysis in Table 3 as the total dollar value bid on loans is positively related to readability and positivity, but inversely related to deception cues (all results are statistically significant at the 1% level; one tail tests). Moreover, the economic magnitude is meaningful. For instance, based on the sample of all listings (first column), a one standard deviation more positive tone listing averages $73.86 more in bids.

A8. Default Robustness: Competing Risk Model

Our primary analysis uses a probit model to examine the relation between default likelihood and the three linguistic features and finds that loans with descriptions that are more readable, more positive, and contain fewer deception cues are less likely to default. In this section we consider a competing risk model (following Duarte et al. (2012)) to examine the relation between default probability and the linguistic measures. Specifically, following these authors, we denote the loan age as the span of time (i.e., installment cycles) over which the loan payment was observed (i.e., until default or pre-payment). Further following the authors, we model the base hazard function as a quadratic function of loan age.

Table A.4 (analogous to the test in the first column of Table 4) reports the results of the competing risk model and shows the results are robust. The second row, for example, indicates that borrowers with more positive tone are less likely to default as hazard rates are less than one.

[Insert Table A.4 about here]

A9. Default, Interest Rate, and 1/(1+*r*)

Iyer et al. (2016, Internet Appendix) show (theoretically) a direct linear relation between default probability and $(1+r)^{-1}$. Therefore, we test whether investors fully incorporate the information contained in the linguistic dimensions by evaluating the relation between default and the linguistic measures when including $(1+r)^{-1}$ as an explanatory variable. In this section, as a robustness test, we repeat this exercise using the borrower rate (r) rather than the inverse of $(1+r)$. Table A.5 reports the marginal effects from a probit regression of default on the borrower rate (r), the three linguistic dimensions, and the full set of control variables. The results are fully consistent with the results in Table 5. First, even when controlling for risk grades (and other characteristics), borrowers that are more likely to default are charged higher rates (i.e., the coefficient on borrower rate is positive and differs significantly from zero at the 1% level). Moreover, consistent with the hypothesis that lenders fail to fully account for the informational content of the linguistic measures, the marginal effects in Table A.5 for the three linguistic characteristics are nearly identical to the values in Table 5. The results suggest that the linguistic measures—especially deception cues—continue to explain default likelihood even when accounting for interest rates (and the other control variables).

[Insert Table A.5 about here]

A10. Default, Interest Rates, and Linguistics: OLS Robustness

Because default is a binary variable, our primary analysis examines the relation between default and the linguistic characteristics controlling for interest rates with a probit regression. The Iyer et al. (2016) model, however, generates a linear relation between default likelihood and $(1+r)^{-1}$. Thus, as a robustness test, we reexamine the relation between default and linguistics controlling for interest rates (and the other control variables) in an OLS regression.

The first column of Table A.6 reports coefficients from an OLS regression of the default indicator on $(1+r)^{-1}$, the three linguistic measures, and the full set of control variables. The second column reports analogous statistics, but replaces $(1+r)^{-1}$ with the borrower rate (r). We continue to find evidence that Prosper investors use information beyond credit grades to price loans, i.e., controlling for credit grades (and the other control variables), borrowers who are more likely to default are charged higher rates. We also continue to find evidence that lenders fail to fully incorporate the informational content of the linguistic measures—especially with respect to deception cues. Moreover, the regression coefficients in Table A.6 are very similar in magnitude to the marginal effects reported in Table 5.

[Insert Table A.6 about here]

A11. Control Variables Coefficient Estimates

To conserve space, our manuscript does not report marginal effects or coefficients for the control variables. In this section, we report the full set of estimates for Tables 3-5 in our paper. Specifically, Table A.7, A.8, and A.9 report the full set of estimates for the analyses in Tables 3, 4, and

5, respectively. Although we do not discuss these results in detail, they are largely consistent with previous work and expectations, e.g., higher credit grade borrowers are more likely to receive funding and are charged lower rates.

[Insert Tables A.7, A.8, and A.9 about here]

A12. Photos and Linguistics

As discussed above and in our study (and Appendix A1 above), Duarte et al. (2012) and Ravina (2019) examine whether the photos that accompany many of the Prosper listings influence which listings get funded and default rates. Specifically, Duarte, Siegel, and Young find borrowers who appear trustworthy, "…the willingness, not the ability, that a potential borrower will repay her loan provided she has the resources to do so…." are more likely to have their listings funded and pay lower rates. Moreover, more trustworthy borrowers are less likely to default. Similar to our analysis of the linguistic dimensions, the authors also find evidence that lenders fail to fully incorporate the soft information into prices, i.e., more trustworthy borrowers' rate, while lower than non-trustworthy borrowers, is still too high given their default probability. Ravina (2019) finds that beautiful borrowers are more likely to have their listings funded but are also more likely to default than average-looking borrowers. In contrast, Ravina finds that Black borrowers are less likely to have their loans funded, but more likely to default.

In this section we examine the possibility that the linguistic dimensions that are the focus of our study are related to characteristics captured by the images. As shown in Tables 3 and 4, our sample consists of 215,930 listings of which 16,044 were funded and became loans. As shown in Table A.10, approximately half the listings contain a photo and of those that had a photo, approximately 65% contained at least one human face while the remaining 35% were photos of non-humans (e.g., pets, home, auto, logo, or humans that were unidentified by the software detailed below). Similarly, of the loans that were funded, approximately 41% included an image with a human face.

[Insert Tables A.10 about here]

We use two automated processes to evaluate the photos. First, we use the Microsoft Face API (https://azure.microsoft.com/en-us/services/cognitive-services/face/#demo) to collect image gender and emotion. If the software does not identify a human face, we classify the photo as non-human (we manually checked a subset of observations and find in all cases, the image is non-human). We define the Female indicator as 1 if the photo contains only females (e.g., the indicator is set to 0 if the photo includes a wife and husband). The Microsoft Face API generates a score of 0 to 1 for eight emotions: happiness, neutral, surprise, anger, contempt, disgust, fear, and sadness. We code the variable positive emotion as 1 if the API codes happiness as the strongest emotion, 0 if the API scores neutral or surprise as the strongest emotion, and -1 if the API scores anger, contempt, disgust, fear, or sadness as the strongest emotion. We then use the Haystack artificial intelligence algorithm (www.haystack.ai) to compute ethnicity and attractiveness scores. Specifically, we code the indicator White race equal to 1 if all individuals in the photo are identified as White race and zero otherwise. The Haystack algorithm also scores attractiveness on a scale from 1 to 10 for each human face. For images that contain more than one face (e.g., wife and husband), we use the average attractiveness.

Table A.11 reports the descriptive statistics for the four image characteristics for our sample. Similar to the linguistic dimensions, there are an infinite number of potential image characteristics. We focus on gender, race, positive emotion, and attractiveness because the technology allows us to measure these characteristics at scale. Clearly, there are other important dimensions that we miss. For instance, although Duarte et al. (2012) find trustworthiness an important image characteristic, we do not attempt to measure perceived trustworthiness (for the 71,873 photos in our data).

[Insert Tables A.11 about here]

One potential concern is that the linguistic dimensions that are the focus of our study are strongly correlated with, and therefore proxy for, image characteristics. Table A.12 reports the correlations between the three linguistic dimensions and the four image characteristics. We find little evidence that the image and linguistic attributes are strongly related. The (absolute) average correlation is 2.6% and the strongest correlation is 5.3%.

[Insert Tables A.12 about here]

We repeat our primary tests for (1) the sample without images, and (2) the same with human images. In the latter case, we include the four image attributes. We begin with a probit of funding success (i.e., directly analogous to Table 3) on linguistic dimensions and loan characteristics for the sample without photos, and on linguistic dimensions, photo dimensions, and loan characteristics for the sample with photos. We include the same set of controls as in Table 3 (i.e., verified hard credit information, unverified credit information, and easily quantifiable nonstandard data). The results for both samples are reported in Table A.13.

[Insert Table A.13 about here]

The results in Table A.13 reveal that, consistent with our primary results, the linguistic dimensions continue to help predict which loans get funded when the sample is limited to listing without photos and when the sample is limited to listings with human images and include the four image characteristics (gender, race, positive emotion, and attractiveness). Specifically, listings that are more readable and have fewer deception cues are more likely to receive funding (statistically significant at the 1% level for both variables in both samples). While the coefficient associated with positive tone has the expected sign, the results are only statistically significant at the 10% level for the sample without images based on a two-tail test. When based on a one-tail test, the coefficient associated with positive tone is statistically significant in both samples. Consistent with Ravina

(2019), the results for the image characteristics suggest that White applicants and more attractive applicants are more likely to have their loans funded.

We next consider how the linguistic dimensions are related to default likelihood when limited to the sample without images and when limited to the sample with human images and including the image characteristics. The first two columns of Table A.14 are analogous to the first column of Table 4 and report the coefficients from a probit regression (default equals one) on the linguistic dimensions and controls (verified hard credit information, unverified credit information, and easily quantifiable nonstandard data) for the sample without images. Consistent with Table 4, the results reveal that less readable texts, more positive tone, and fewer deception cues are all associated with a lower default likelihood. The results in the second column add the image characteristics to the regression for the sample that contains human images. The results reveal that less positive tone and more deception cues continue to predict higher default likelihood when including the four image characteristics for the sample of loans that include human images. Although the coefficient associated with readability has the expected sign, the coefficient does not differ meaningfully from zero.

[Insert Table A.14 about here]

Consistent with the analysis reported in Table 5, the results in the final two columns of Table A.14 reveal that even after accounting for the rate charged, loans with greater readability, greater positivity, and fewer deception cues are less likely to default for the sample without photos. For the sample with photos, the coefficients associated with positivity and deception cues maintain the expected signs and are statistically significant at the 1% level. Although the coefficient associated with readability has the expected sign, it is no longer statistically significant. In short, the results in the last two columns reveal that Prosper investors fail to fully account for the information contained in the linguistic dimensions.

In sum, the results reported in Tables A.10-A.14 reveal no evidence that the photos and linguistic characteristics meaningfully overlap. Our conclusions remain largely unchanged when limiting the sample to listings without images or when limiting the sample to listings with human images and accounting for gender, race, emotion, and attractiveness.

References

Asay, H.; W. Scott; B. Elliott; and K. M. Rennekamp. "Disclosure Readability and the Sensitivity of Investors' Valuation Judgments to Outside Information." *Accounting Review,* 92 (2017), 1-25.

Bonsall, S. B.; A. J. Leone; B. P. Miller; and K. Rennekamp. "A Plain English Measure of Financial Reporting Readability." *Journal of Accounting and Economics,* 63 (2017), 329-357.

Crammer, K., and Y. Singer. "On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines." *Journal of Machine Learning Research,* 2 (2001), 265-292.

Dorfleitner, G.; C. Priberny; S. Schuster; J. Stoiber; M. Weber; I. de Castro; and J. Kammler. "Description-text Related Soft Information in Peer-to-Peer Lending–Evidence from Two Leading European Platforms." *Journal Banking and Finance,* 64 (2016), 169-187.

Dougal, C.; J. Engelberg; D. Garcia; and C. Parsons. "Journalists and the Stock Market." *Review of Financial Studies*, 25 (2012), 639-679.

DuBay, W. H. *The Principles of Readability*. Costa Mesa, CA: Impact Information (2004).

Duarte, J.; S. Siegel; and L. Young. "Trust and Credit: The Role of Appearance in Peer-to-peer Lending." *Review of Financial Studies,* 25 (2012), 2455-2484.

Ehsan, N., and H. Faili. "Grammatical and Context-Sensitive Error Correction using a Statistical Machine Translation Framework." *Software: Practice and Experience,* 43 (2013), 187-206.

Franco, G.; O. Hope; D. Vyas; and Y. Zhou. "Analyst Report Readability." *Contemporary Accounting Research,* 32 (2015), 76-104.

Gunning, R. "The Fog Index After Twenty Years." *Journal of Business Communication,* 6 (1969), 3-13.

Healy, P. M., and K. G. Palepu. "Information Asymmetry, Corporate Disclosure, and the Capital Markets: A Review of the Empirical Disclosure Literature." *Journal of Accounting and Economics,* 31 (2001), 405-440.

Herzenstein, M.; S. Sonenshein; and U. M. Dholakia. "Tell Me a Good Story and I May Lend You Money: The Role of Narratives in Peer-to-Peer Lending Decisions." *Journal of Marketing Research,* 48 (2011), S138-S149.

Huang, A. H.; A. Y. Zang; and R. Zheng. "Evidence on the Information Content of Text in Analyst Reports." *Accounting Review,* 89 (2014), 2151-2180.

Hwang, B., and H. H. Kim. "It Pays to Write Well." *Journal of Financial Economics,* 124 (2017), 373-394.

Iyer, R.; A. I. Khwaja; E. F. P. Luttmer; and K. Shue. "Screening Peers Softly: Inferring the Quality of Small Borrowers." *Management Science,* 62 (2016), 1554-1577.

Joachims, T.; T. Finley; and C. Yu. “Cutting-plane Training of Structural SVMs.” *Machine Learning,* 77 (2009), 27-59.

Jurafsky, D., and J. H. Martin. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice Hall (2000).

Lawrence, A. “Individual Investors and Financial Disclosure.” *Journal of Accounting and Economics,* 56 (2013), 130-147.

Lehavy, R.; F. Li; and K. Merkley. “The Effect of Annual Report Readability on Analyst Following and the Properties of their Earnings Forecasts.” *Accounting Review,* 86 (2011), 1087-1115.

Li, F. “Annual Report Readability, Current Earnings, and Earnings Persistence.” *Journal of Accounting and Economics,* 45 (2008), 221-247.

Liskovich, I., and M. Shaton. “Borrowers in Search of Feedback: Evidence from Consumer Credit Markets.” Working paper, University of Texas Austin (2017).

Lo, K.; F. Ramos; and R. Rogo. “Earnings Management and Annual Report Readability.” *Journal of Accounting and Economics,* 63 (2017), 1-25.

Loughran, T., and B. McDonald. “When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *Journal of Finance,* 66 (2011), 35-65.

Loughran, T., and B. McDonald. “Measuring Readability in Financial Disclosures.” *Journal of Finance*, 69 (2014), 1643-1671.

Michels, J. “Do Unverifiable Disclosures Matter? Evidence from Peer-to-Peer Lending.” *Accounting Review,* 87 (2012), 1385-1413.

Miłkowski, M. “Developing an Open-Source, Rule-Based Proofreading Tool.” *Software: Practice and Experience,* 40 (2010), 543-566.

Miller, B. P. “The Effects of Reporting Complexity on Small and Large Investor Trading.” *Accounting Review,* 85 (2010), 2107-2143.

Mitton, R. “Spelling Checkers, Spelling Correctors and the Misspellings of Poor Spellers.” *Information Processing & Management*, 23 (1987), 495-505.

Naber, D. “A Rule-Based Style and Grammar Checker.” (2003) Available: http:// www.danielnaber.de/languagetool/download/style_and_grammar_checker.pdf.

Netzer O.; A. Lemaire; and M. Herzenstein. “When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications.” *Journal of Marketing Research,* 56 (2019), 960-980.

Ravina, E. “Love & Loans: The Effect of Beauty and Personal Characteristics in Credit Markets.” Working paper, Northwestern University (2019).

Table A.1

Correlation between Linguistic Measures

This table reports the correlation between the three linguistic measures. The variables are described in Table 1 of the manuscript. This appendix provides construction details. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively. The sample includes 215,931 listings between 02/12/2007 and 10/16/2008.

	Readability	Positivity	Deception Cues
READABILITY	1.000		
TONE	-0.118***	1.000	
DECEPTIOIN_CUES	0.039***	-0.130***	1.000

Table A.2
Linguistic Measures and Funding Success—Open for Duration Listings Only

The first column reports marginal effects (standard errors in parentheses) from a probit regression of the funding success indicator on hard credit information, unverified credit information, easily quantifiable nonstandard information, and three linguistic features—readability, positivity, and deception cues. The second through fourth columns report marginal effects (standard errors in parentheses) from Tobit regressions of number of bids, total dollar amount bid divided by amount requested, and interest rate, respectively, on the same independent variables. Table 1 provides definitions for all variables. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels from two-tailed tests, respectively. The sample includes 164,098 listings and 12,874 loans between 02/12/2007 and 10/16/2008.

	Loan Funded Indicator	Number of Bids	Total $Amount Bid / $Amt. Requested	Interest Rate
READABILITY	0.0024***	1.119***	0.0179***	-0.0485
	(0.0004)	(0.242)	(0.0035)	(0.0319)
TONE	0.0016***	1.460***	0.0178***	-0.155***
	(0.0006)	(0.390)	(0.0056)	(0.0451)
DECEPTIOIN_CUES	-0.0020***	-3.630***	-0.0233***	0.0466
	(0.0004)	(0.241)	(0.0034)	(0.0322)
Observations	164,098	164,098	164,098	12,874
Verified Hard Credit Information	YES	YES	YES	YES
Unverified Credit Information	YES	YES	YES	YES
Easily Quant. Nonstandard Data	YES	YES	YES	YES

Table A.3
Linguistic Measures and Funding Success—Total Dollars Bid

The first column reports marginal effects (standard errors in parentheses) from a Tobit regression of the total dollars bid on hard credit information, unverified credit information, easily quantifiable nonstandard information, and three linguistic features—readability, positivity, and deception cues. The second column reports analogous results when the sample is limited to "open for duration" listings. Table 1 provides definitions for all variables. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels from one tailed test, respectively. The sample period is 02/12/2007 to 10/16/2008.

	Total $Amount Bid (all listings)	Total $Amount Bid (open for duration only)
READABILITY	82.32***	96.26***
	(16.70)	(21.11)
TONE	73.86***	105.00***
	(27.17)	(33.92)
DECEPTIOIN_CUES	-277.10***	-312.50***
	(16.65)	(20.97)
Observations	215,930	164,098
Verified Hard Credit Information	YES	YES
Unverified Credit Information	YES	YES
Easily Quant. Nonstandard Data	YES	YES

Table A.4
Robustness Tests: Competing Risk Model—Linguistic Style and Default Risk

This table reports sub-distribution hazard ratios from a competing risk model of default or prepayment on hard credit information, unverified credit information, easily quantifiable nonstandard information, and three linguistic features—readability, positivity, and deception cues. Table 1 provides variable definitions. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels from one tailed test, respectively (results are based on tests of difference from one). The sample includes 16,044 funded loans between 02/12/2007 and 10/16/2008.

	Competing Risk Model
READABILITY	0.977*
	(0.0134)
TONE	0.947***
	(0.0190)
DECEPTIOIN_CUES	1.062***
	(0.0148)
Observations	16,044
Verified Hard Credit Information	YES
Unverified Credit Information	YES
Easily Quant. Nonstandard Data	YES

Table A.5
Do Lenders Fully Account for the Informational Content of Borrower's Writing? Robustness using *r* instead of $(1+r)^{-1}$

This table reports marginal effects (standard errors in parentheses) from a probit regression of the default indicator on hard credit information, unverified credit information, easily quantifiable nonstandard data, the borrower interest rate (*r*) in percent, and three linguistic features—readability, positivity, and deception cues. Table 1 provides variable definitions. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels from two tailed test, respectively. The sample includes 16,044 funded loans between 02/12/2007 and 10/16/2008.

	Default Indicator
BWR_IR (*r*)	0.0119***
	(0.0013)
READABILITY	-0.0074*
	(0.0044)
TONE	-0.0157**
	(0.0064)
DECEPTIOIN_CUES	0.0212***
	(0.0045)
Observations	16,044
Verified Hard Credit Information	YES
Unverified Credit Information	YES
Easily Quant. Nonstandard Data	YES

Table A.6
Do Lenders Fully Account for the Informational Content of Borrower's Writing? OLS Robustness

This table reports coefficients from an OLS regression of the default indicator on hard credit information, unverified credit information, easily quantifiable nonstandard data, (1+borrower interest rate)$^{-1}$, and three linguistic features—readability, positivity, and deception cues. The second column reports results when replacing (1+*r*)$^{-1}$ with the borrower interest rate (*r*) in percent. Table 1 provides variable definitions. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels from two tailed test, respectively. The sample includes 16,044 funded loans between 02/12/2007 and 10/16/2008.

	Default Indicator	Default Indicator
1/(1+BWR_IR)	-2.066***	
	(0.181)	
BWR_IR (*r*)		0.0118***
		(0.0012)
READABILITY	-0.0067*	-0.0068*
	(0.0040)	(0.0040)
TONE	-0.0130**	-0.0138**
	(0.0058)	(0.0058)
DECEPTIOIN_CUES	0.0181***	0.0184***
	(0.0040)	(0.0040)
Observations	16,044	16,044
Verified Hard Credit Information	YES	YES
Unverified Credit Information	YES	YES
Easily Quant. Nonstandard Data	YES	YES

Table A.7
Linguistic Measures and Funding Success (Table 3 full set of coefficients)

	Loan Funded Indicator	Number of Bids	Total $Amount Bid / $Amt. Requested	Interest Rate
READABILITY	0.0024***	0.945***	0.0157***	-0.0375
TONE	0.0009*	0.968***	0.0100**	-0.112***
DECEPTIOIN_CUES	-0.0020***	-3.146***	-0.0211***	0.0342
CR_GRADE_2(A)_IND	-0.122***	-26.74***	-0.398***	0.424***
CR_GRADE_3(B)_IND	-0.176***	-42.05***	-0.627***	0.746***
CR_GRADE_4(C)_IND	-0.241***	-66.39***	-0.981***	0.849***
CR_GRADE_5(D)_IND	-0.285***	-87.87***	-1.301***	1.700***
CR_GRADE_6(E)_IND	-0.311***	-107.6***	-1.586***	3.547***
CR_GRADE_7(HR)_IND	-0.320***	-121.5***	-1.794***	3.627***
BANKCARD_UTIL	0.0023**	0.473	0.0158*	0.351***
AMT_DELQ	-0.00418***	-2.156***	-0.0322***	0.0689***
DEQL(LAST_7_YR)	-0.0023***	-0.857***	-0.0127***	0.165***
INQ_LAST_6_MTHS	-0.0021***	-0.871***	-0.0127***	0.0556***
PUB_REC(10_YR)	-0.00203***	-1.036***	-0.0143***	0.0551
PUB_REC(LAST_YR)	0.0029*	1.300	0.0224*	-0.0194
CURRENT_LOC	0.00525***	-0.579	-0.0114	-0.171
OPEN_LOC	-0.0130***	-6.223***	-0.0847***	0.315**
CR_HIS(MTHS)	-0.0031***	-1.277***	-0.0182***	0.0705**
REVOLV_CR_BAL	-0.0006***	-0.0662	-0.001	-0.0248*
LOAN_AMT	-0.0249***	3.317***	-0.226***	0.0173
DTI	-0.0059***	-2.398***	-0.0361***	0.0803***
HOMEOWNER_IND	0.00245***	3.973***	0.0483***	0.0038
INC. IND_2 ($25k to <$50k)	0.0180***	17.280***	0.264***	0.0479
INC. IND_3 ($50k to <$75k)	0.0232***	23.310***	0.344***	-0.0875
INC. IND_4 ($75k to <$100k)	0.0280***	28.010***	0.411***	-0.0905
INC. IND_5 (>$100k)	0.0306***	30.930***	0.446***	-0.132
INC. IND_6 (missing)	0.0270***	33.360***	0.448***	-0.147
FULL-TIME_EMPL_IND	-0.0299	-15.380	-0.0201	-1.958*
NOT_EMPLOYED_IND	-0.0380*	-21.10**	-0.108	-2.572**
PART-TIME_EMPL_IND	-0.0342	-21.19**	-0.0974	-2.094**
RETIRED_IND	-0.0406*	-19.99*	-0.0789	-2.182**
SELF-EMPLOYED_IND	-0.0489**	-28.61***	-0.224	-1.941*
BWR_MAX_IR	0.274***	232.2***	3.366***	69.33***
OPEN_FOR_DUR	-0.0198***	3.965***	0.0783***	-4.043***
PHOTO_IND	0.0104***	6.268***	0.0942***	-0.147***
ENDT_IND	0.0062***	6.879***	0.120***	-0.153*
FRIENDS_IND	0.0094***	3.359***	0.0364***	-0.0788
GROUP_MEM	0.0300***	23.49***	0.361***	-0.1180**
TEXT_LEN	0.0034***	-1.688***	0.0030	0.0801**
Observations	215,930	215,930	215,930	16,044

Table A.8
Linguistic Measures and Default Risk (Table 4 full set of coefficients)

	Default Indicator	%Principal Repaid
READABILITY	-0.0079*	1.797**
TONE	-0.0171***	2.736**
DECEPTIOIN_CUES	0.0216***	-4.095***
CR_GRADE_2(A)_IND	0.0849***	-17.82***
CR_GRADE_3(B)_IND	0.130***	-25.17***
CR_GRADE_4(C)_IND	0.102***	-19.45***
CR_GRADE_5(D)_IND	0.0983***	-18.24***
CR_GRADE_6(E)_IND	0.0870***	-16.54***
CR_GRADE_7(HR)_IND	0.153***	-26.92***
BANKCARD_UTIL	0.0619***	-9.975***
AMT_DELQ	0.00828***	-1.695***
DEQL(LAST_7_YR)	-0.0135***	3.260***
INQ_LAST_6_MTHS	0.0205***	-3.724***
PUB_REC(10_YR)	0.0198***	-3.244***
PUB_REC(LAST_YR)	-0.0386*	8.933**
CURRENT_LOC	-0.00403	3.861
OPEN_LOC	0.00707	-3.274
CR_HIS(MTHS)	0.00552	-0.549
REVOLV_CR_BAL	-0.0122***	2.236***
LOAN_AMT	-0.0074	0.745
DTI	0.0175***	-2.629***
HOMEOWNER_IND	0.0827***	-14.58***
INC. IND_2 ($25k to <$50k)	0.0391	-10.11
INC. IND_3 ($50k to <$75k)	0.0318	-11.43
INC. IND_4 ($75k to <$100k)	0.000263	-4.784
INC. IND_5 (>$100k)	0.00759	-4.586
INC. IND_6 (missing)	-0.00642	-4.422
FULL-TIME_EMPL_IND	0.0138	-25.92
NOT_EMPLOYED_IND	0.0531	-37.04
PART-TIME_EMPL_IND	-0.0358	-13.61
RETIRED_IND	0.0323	-33.75
SELF-EMPLOYED_IND	0.112	-44.98
BWR_MAX_IR	1.200***	-258.7***
OPEN_FOR_DUR	-0.0754***	15.89***
PHOTO_IND	-0.0126	2.728*
ENDT_IND	0.0649***	-13.04***
FRIENDS_IND	-0.0501***	11.45***
GROUP_MEM	0.0219**	-4.473**
TEXT_LEN	0.0127**	-2.706**
Observations	16,044	16,044

Table A.9
Do Lenders Fully Account for the Informational Content of Borrower's Writing? (Table 5 full set of coefficients)

	Default Indicator	%Principal Repaid
1/(1+BWR_IR)	-2.158***	419.4***
READABILITY	-0.0073*	1.677**
TONE	-0.0149**	2.249*
DECEPTIOIN_CUES	0.0209***	-3.943***
CR_GRADE_2(A)_IND	0.0749***	-15.33***
CR_GRADE_3(B)_IND	0.111***	-21.01***
CR_GRADE_4(C)_IND	0.0809***	-14.76***
CR_GRADE_5(D)_IND	0.0644***	-11.05***
CR_GRADE_6(E)_IND	0.0301	-4.983
CR_GRADE_7(HR)_IND	0.0946***	-15.18***
BANKCARD_UTIL	0.0559***	-8.839***
AMT_DELQ	0.0073***	-1.493***
DEQL(LAST_7_YR)	-0.0159***	3.683***
INQ_LAST_6_MTHS	0.0197***	-3.547***
PUB_REC(10_YR)	0.0190***	-3.075***
PUB_REC(LAST_YR)	-0.0390*	9.064**
CURRENT_LOC	-0.0024	3.593
OPEN_LOC	0.0035	-2.611
CR_HIS(MTHS)	0.0043	-0.324
REVOLV_CR_BAL	-0.0119***	2.148***
LOAN_AMT	-0.0078	0.804
DTI	0.0159***	-2.356***
HOMEOWNER_IND	0.0829***	-14.50***
INC. IND_2 ($25k to <$50k)	0.0365	-9.623
INC. IND_3 ($50k to <$75k)	0.0314	-11.31
INC. IND_4 ($75k to <$100k)	-0.0004	-4.693
INC. IND_5 (>$100k)	0.0072	-4.451
INC. IND_6 (missing)	-0.0070	-4.235
FULL-TIME_EMPL_IND	0.0426	-29.61
NOT_EMPLOYED_IND	0.0884	-41.8
PART-TIME_EMPL_IND	-0.0048	-17.68
RETIRED_IND	0.0653	-38.32
SELF-EMPLOYED_IND	0.140	-48.44
BWR_MAX_IR	0.168	-54.38**
OPEN_FOR_DUR	-0.0162	4.404**
PHOTO_IND	-0.0103	2.27
ENDT_IND	0.0672***	-13.50***
FRIENDS_IND	-0.0497***	11.44***
GROUP_MEM	0.0228**	-4.679***
TEXT_LEN	0.0107*	-2.288**
Observations	16,044	16,044

Table A.10
Sample Sizes with and without Photos

This table reports the number of listings and loans with photos as well as the number with photos that include human images.

	Listings	Loans
Photos with human faces	71,873	6,581
Photos without human faces	39,681	3,755
Total with photos (a)	111,554	10,336
Total without photo (b)	104,376	5,708
Total (a)+(b)	215,930	16,044

Table A.11
Descriptive Statistics for Images

We use two automated processes to compute image characteristics for 71,873 listings that contain a human image. The Microsoft Face API computes gender and emotion. The Haystack artificial intelligence algorithm computes race and attractiveness. Female is equal to 1 if the image only contains female faces. White is equal to 1 if the image only contains White faces. Positive emotion equals 1 if happiness is identified as the strongest emotion, 0 if neutral or surprise is the strongest emotion, and -1 if anger, contempt, disgust, fear, or sadness is the strongest emotion. Attractiveness is based on a scale of 1 to 10.

Variable	N	Mean	Standard deviation	Minimum	Maximum
FEMALE	71,873	0.4152	0.4928	0	1
WHITE_RACE	71,873	0.5845	0.4928	0	1
POSITIVE_EMOTION	71,873	0.5908	0.6599	-1	1
ATTRACTIVENESS	71,873	6.2298	1.6704	4	10

Table A.12
Correlation between Linguistic and Image Attributes

This table reports the correlation between the four image characteristics and the three linguistic characteristics. The variables are described in Table A.11. The sample includes 71,873 listings containing human images between 02/12/2007 and 10/16/2008.

Correlation between linguistic and image dimensions (*n*=71,873)			
	Readability	Positivity	Deception Cues
FEMALE	0.0127	0.0227	-0.0423
WHITE_RACE	0.0210	0.0532	-0.0016
POSITIVE_EMOTION	0.0294	0.0412	-0.0228
ATTRACTIVENESS	0.0094	0.0397	-0.0201

Table A.13
Funding Success, Linguistic Attributes, and Image Attributes

The first column reports marginal effects from a probit regression of the funding success indicator on controls (hard credit information, unverified credit information, easily quantifiable nonstandard information) and three linguistic features—readability, positivity, and deception cues for the sample of listings that do not contain an image. The second column reports an analogous regression for the sample of listings that include a human image and adds the four imagine characteristics (gender, race, positive emotion, and attractiveness) as regressors. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels from two-tailed tests, respectively.

	Listings with no image	Listings with human image
READABILITY	0.0009***	0.0041***
TONE	0.0008*	0.0017
DECEPTIOIN_CUES	-0.0014***	-0.0031***
FEMALE		-0.0013
WHITE_RACE		0.0119***
POSITIVE_EMOTION		0.0008
ATTRACTIVENESS		0.0028***
Controls	YES	YES
Observations	104,376	71,873

Table A.14
Default Analysis, Linguistic Attributes, and Image Attributes

The first column reports marginal effects (standard errors in parentheses) from a probit regression of the default indicator on controls (hard credit information, unverified credit information, easily quantifiable nonstandard information) and three linguistic features—readability, positivity, and deception cues for the sample of funded loans that do not contain an image. The second column reports an analogous regression for the sample of listings that include a human image and adds the four imagine characteristics (gender, race, positive emotion, and attractiveness) as regressors. The final two columns repeat both sets of regression but add the inverse interest as an additional regressor. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels from two-tailed tests, respectively.

	Default Indicator		Default Indicator (with inverse interest rate)	
	No Image	Image	No Image	Image
1/(1+BWR_IR)			1.9613***	2.1479***
READABILITY	-0.0181***	-0.0071	-0.0173***	-0.0065
TONE	-0.0232**	-0.0289***	-0.0205*	-0.0270***
DECEPTIOIN_CUES	0.0195***	0.0272***	0.0190***	0.0259***
FEMALE		0.0561***		0.0554***
WHITE_RACE		-0.0134		-0.0125
POSITIVE_EMOTION		0.0014		0.0016
ATTRACTIVENESS		0.0474***		0.0417***
Controls	YES	YES	YES	YES
Observations	5,708	6,581	5,708	6,581