Supporting Information for

Even Lawyers Don't Like Legalese

Eric Martínez, Francis Mollica and Edward Gibson

Eric Martínez.
E-mail: ericmartmit.edu

This PDF file includes:

Supporting Information Text

1. Experiment 1

A. Author Recognition Test Analyses. As noted in the main text, for Experiment 1 we administered the Author Recognition Test (a validated proxy for reading ability) to participants as part of each trial. Although this was mainly used as a filler task, and we made no pre-registered predictions regarding the results, for transparency we report the results of its potential effect on comprehension and recall data. The results of all of these analyses were convergent with the findings in Martinez, Mollica & Gibson's sample of laypeople (1).

When adding ART score as a fixed-effect predictor to our comprehension model, we found a main effect of ART score on comprehension (β =.221,SE=.078,p=.005), such that those who had higher ART scores had higher comprehension scores. However, we did not find a significant interaction between ART and register (p=.075). When adding ART score as a fixed-effect predictor to our recall model, we did not find a main effect of ART score on recall (p=.787).

B. Recall annotation details. Two trained research assistants coded whether a proposition was successfully recalled, using the same method as (1). In particular, they were given a participant's retelling of a passage and then asked whether each legally relevant proposition of the passage was (a) fully recalled; (b) partially recalled; or (c) not recalled. Coders were told that for a response to count as "fully recalled," it did not have to be recalled verbatim (i.e. they can use their own words or syntax), so long as they were confident that the meaning of what subject wrote is the same as the proposition.

For example, suppose the original text said "A court in Boston will resolve the dispute," and the participant wrote "something will be resolved by a court." When coding responses, a coder might see three propositions that say: (i) "A court in Boston," (ii) "will resolve," and (iii) "the dispute."

For (i), the coder would put a 0.5 for "partially recalled" (since "in Boston" was missing from "a court"); for (ii), the coder would put a 1 for "fully recalled" (since "will be resolved" means basically the same thing as "will resolve"), and for (iii), the coder would put a 0 for "not recalled" (since "the dispute" was not in the response).

To reduce potential bias, coders were unaware of whether a participant had seen or recalled the simple or legalese version of a text. Moreover, the rubric that the coders used to score participants' responses was the legalese version: for each proposition, they were given the language of that proposition in legalese, and were told to score a participants' response as having recalled that proposition if it had language that had the same meaning as that proposition. Thus, any differences in recall favoring the Plain English version would arise in spite of the coding bias, which was towards the Legalese version.

Of the roughly 650 retellings within the lawyer data, one coder was responsible for coding 100 percent of the retellings, while the second coder was responsible for coding a random subset (20%) of the retellings, so as to assess inter-rater reliability. Coder reliability was assessed with Cohen's kappa coefficient (2, 3).

We adjudicated ties as follows: (i) a tie between one "fully recalled" judgment and one "not recalled" judgment resulted in a final "partially recalled" judgment; (ii) a tie between one "fully recalled" judgment and one "partially recalled" judgment resulted in a final "fully recalled" judgment for a given proposition; and (iii) a tie between one "partially recalled" judgment and one "not recalled" judgment resulted in a final "not recalled" judgment. For our regression analyses, we perform both a conservative analysis (recoding "partially recalled" as "not recalled') and an anti-conservative analysis (recoding "partially recalled" as "fully recalled"). Our results do not qualitatively change, so we will only report the conservative analysis in the main writeup.

C. Anti-conservative recall analysis. As noted in the main writeup, for our recall analyses, we performed both a conservative analysis and an anti-conservative analysis. For the conservative analysis, As our results did not qualitatively change, so we only reported the conservative analysis in text.

For both the conservative and anti-conservative analyses, we conducted a mixed effect logistic regression with register, legal training and the interaction between the two as fixed effects, and participant as random effects, with register as a random slope for each. As in the conservative analysis, for the anti-conservative analysis, we found a main effect of register ($\beta = .225$, SE=.098, $p = .022$) and legal training ($\beta = -.340$, SE=.153, $p = .026$) on recall. As in the conservative analysis, we did not find a main effect of the interaction between register and condition ($p = .174$).

D. Subjective rating analyses. As noted in the main writeup, in addition to recall and comprehension analyses, we also asked participants to rate how difficult a text they found the text (a) for themselves; (b) for the average layperson; and (c) for the average lawyer.

Below is the wording of each of the prompts:

- (you): "How complex/difficult do you find this text to understand?"
- (lawyer):"How complex/difficult do you think the average layperson/non-lawyer would find this text to understand?"
- (layperson): "How complex/difficult do you think the average lawyer would find this text to understand?"

The 5 answer choices for each prompt were as follows:

- extremely simple/easy

- somewhat simple/easy
- neither complex/difficult nor simple/easy
- somewhat complex/difficult
- extremely complex/difficult

Results are visualized in the main text.

To analyze these results, we ran three different models.

First, we conducted a model that compared how difficult lawyers and laypeople predicted texts would be for the average layperson.

Second, we conducted a model that compared (a) how difficult lawyers predicted texts would be for the average layperson compared with (b) how difficult lay participants perceived the texts to be for themselves.

Third, we conducted a model that compared (a) how difficult laypeople predicted texts would be for the average lawyer compared with (b) how difficult lay participants perceived the texts to be for themselves.

This model's predictions are less relevant for the curse of knowledge hypothesis, but for robustness purposes we report it anyway.

The model for all three models was as follows:

- clmm(as.factor(Response) condition*training + (1 + condition | subject) + (1 + condition | item), data =.,)

The only difference among the three models was that the "response" variable was filtered to include a different subset of the data (to include the relevant conditions)

For the first model, we found a main effect of condition (β =-1.784, SE=.161, p<.001), but not training (p=.974), nor the interaction between condition and training (p=.127).

That is, contrary to the predictions of the curse of knowledge hypothesis, we did not find evidence that lawyers underestimated the difficulty of legal texts for non-lawyers, nor did they particularly underestimate the difficulty of legal texts written in legalese.

For the second model, we found a main effect of condition (β =-1.784, SE=.161, p<.001). We also found an effect of training (β =-.938, SE=.293, p=.001), with laypeople's own subjective ratings being significantly easier than lawyers' predictions of the average layperson's subjective ratings.

We also found an interaction between training and condition (β =.562, SE=.250, p=.025), such that laypeople's ratings of simple texts were disproportionately high relatively to lawyers' predictions of those texts relative to the groups' legalese ratings.

For the third model, we found a main effect of condition (β =-2.235, SE=-8.718, p<.0001). We also found an effect of training (β =-3.094, SE=.339, p<.0001), with laypeople's predictions of lawyers' ratings being significantly easier than lawyers' subjective ratings of the texts.

We also found an interaction between training and condition (β =.792, SE=.266, p=.003), such that laypeople's predictions of lawyers' ratings of simple texts were disproportionately high relatively to lawyers' ratings of those texts relative to the legalese ratings.

2. Experiment 2

Hypotheses and Predictions. In Experiment 2, we aimed to test the following hypotheses and associated predictions. All of these were pre-registered on OSF.

Hypothesis I: Lawyers simply write in a complex register out of "habit, laziness" (4) or "tradition" (5); they "copy and paste" (Adams, 2022) from existing templates with old, complicated terms because that's the "quickest and cheapest way to produce a contract" (6), not out of any preference.

- Prediction 1: Lawyers will rate plain English contracts as of equal quality as legalese contracts.
- Prediction 2: Lawyers will agree to sign off on a contract written in Plain English.

Hypothesis II: Lawyers write in legalese in order to be accepted by peers. The legalese signals in-group membership (5).

- Prediction 1: Lawyers will rate contracts written in legalese as sounding more "lawyerly" (more appropriate/suitable for a lawyer) than those written in plain English.
- Prediction 2: Lawyers will rate authors of contracts written in legalese more hirable than authors of contracts written in plain English.

Hypothesis III: Lawyers write in legalese as a way of "preserving their monopoly" (7) on legal services and "justifying fees" (5)

- Prediction: Lawyers will predict contracts written in legalese as being more likely to be signed by clients than contracts written in plain English.

Hypothesis IV: Contractual language needs to be complex in order to convey complex legal concepts in a way that "is far more precise than ordinary language" (4) and/or to be enforceable

- Prediction: Plain English contracts will be rated as unenforceable by legal experts

Materials. To evaluate our predictions, we measured two sets of outcome variables. The first set of outcome variables were measured individually for each text and were as follows:

- text quality ("How would you rate the overall quality of the above contract excerpt?")
- enforceability ("Suppose two parties signed a contract that included the above excerpt. Would the excerpt likely be legally enforceable (assuming the rest of the contract was enforceable)?")
- useability ("Suppose someone at your firm drafted a contract that included the above excerpt. Would you and your firm agree to execute it as currently written (assuming the rest of the contract was okay)?")
- hireability ("Suppose the excerpt was drafted by someone outside your firm. Would your firm be likely to hire them to draft future contracts, all else equal?")
- lawyerliness ("Does the style/tone of the excerpt sound appropriate for a lawyer?")
- likelihood of being signed ("Suppose you drafted this excerpt for a client as part of a larger contract. Would a client be likely to sign this contract (assuming the rest of the contract was written in a similar style)?")

Text quality was measured on a scale of 1-5 (1 being "extremely low-quality" and 5 being "extremely high-quality"). All other outcome variables in this set were measured on a yes-no scale.

The second set of outcome variables were measured for each contract pair as opposed to each individual contract. These variables were as follows:

- more usable ("Which of the two versions would you be more likely to execute, given the choice?")
- more likely to be signed ("Which of the two versions would a client be more likely to agree to sign?")
- more lawyerly ("Which of the two versions sounds more appropriate for a lawyer?")
- more hireable ("Suppose the two versions were drafted by two different authors. Which of the two would your firm be more likely to hire to draft future contracts, all else equal?")

All outcome variables in this set were measured on a two-point scale (version 1 or version 2).

A. Analysis Plan. To evaluate Hypothesis I, we conducted two regressions.

- An ordinal regression with the following syntax: `clmm(text quality ~ condition + (1 + condition | item) + (1 + condition | subject), data = .)`
- A logistic regression with the following syntax: `glmer(is usable ~ condition + (1 + condition | item) + (1 + condition | subject), data = ., family = binomial(link = "logit"))`

For Hypothesis II we we conducted exact binomial tests for the more lawyerly and more hireable variables.

For Hypothesis III we conducted the following logistic regression, as well as an exact binomial test:

- `glmer(client would sign ~ condition + (1 + condition | item) + (1 + condition | subject), data = ., family = binomial(link = "logit"))`

For Hypothesis IV, we conducted the following regression:

- `glmer(is enforceable ~ condition + (1 + condition | item) + (1 + condition | subject), data = ., family = binomial(link = "logit"))`

In our pre-registration, we stated that if we encountered issues fitting models, we would use Bayesian regression techniques with similar syntax. We did not encounter issues fitting models, and therefore will report our pre-registered models.

B. Supplementary Results. Results are visualized in Table S1 and Figure S1. As noted in the main text, all of the predictions of hypotheses 1-3 were disconfirmed, and all of the predictions of hypothesis 4 were confirmed.

With regard to hypothesis 1, contrary to the first prediction, we found that lawyers were more likely to say that they would use simple contracts over legalese contracts (β =1.432, SE=.270, p<.001), and rated simple texts as higher quality than legalese texts (β =1.705,SE=.329,p<.001).

With regard to hypothesis 2, contrary to both predictions, participants were more likely to rate the authors of simple texts as hireable compared to the authors of legalese texts ($\beta = 1.835, SE = .318, p < .001$), and we did not find participants to be more likely to rate legalese texts as sounding more lawyerly than simple texts (p=.692).

With regard to hypothesis 3, contrary to the pre-registered prediction, we found that participants were more likely to predict that clients would sign a simple contract compared to a legalese contract ($\beta = 1.232, SE = .338, p < .001$).

With regard to hypothesis 4, in line with the predictions, we found that participants were more likely to say that they would agree to use the simple contracts as written (β =1.705,SE=.329,p<.001), and we did not find a significant difference in how enforceable the different contracts were rated as (p=.717).

C. Exploratory Demographic Analyses. Our main analyses, predictions, and conclusions drawn on the basis of our analyses were limited to those that we included in our pre-registration on OSF.

Although our pre-registered statistical models did not include any demographic analyses, one might wonder whether our results may have been driven by the demographic composition of our lawyer sample.

Below is a description of each of these results as applied to our hypotheses. To help lend a visual sense of the robustness of the results, results for Experiment 2 limited to "experienced" attorneys (those with 10 or more years of practice experience) are visualized in Figure S4. Results limited to "fancy" attorneys (those who attended a top-25 law school or work at a top-200 law firm) are visualized in Figure S5.

To more rigorously account for the possibility of results being driven by demographic factors, we conducted additional versions of our pre-registered analyses, adding each of our demographic variables as fixed-effect predictors. Doing so did not alter our results.

Specifically, in cases where our pre-registered models found a main effect of a given predictor variable, the same was true when adding all of the demographic variables as additional fixed-effect predictors.

Conversely, in cases where our pre-registered models did not find a main effect of a given predictor variable, the same was true when adding all of the demographic variables as additional fixed-effect predictors.

C.1. Curse of Knowledge. With regard to the curse-of-knowledge hypothesis, We added the demographic factors to a modified model of comprehension and recall models that (a) were limited to lawyers; and (b) did not contain fixed-effects related to legal training. These models were as follows:

- `glmer( comprehension ~ condition + age + is fancy + gender + ethnicity + practice experience + (1 + condition | item) + (1 + condition | subject), data = ., family = binomial(link = "logit"))`
- `glmer( recall ~ condition + age + is fancy + gender + ethnicity + practice experience + (1 + condition | item) + (1 + condition | subject), data = ., family = binomial(link = "logit"))`

As with our pre-registered models, these models revealed a main effect of register, such that lawyers were significantly more likely to comprehend (β =.373,SE=.094,p<.0001) and recall (β =.376,SE=.132,p<.0001) legal content written in simple contracts relative to legalese contracts.

The comprehension model revealed a main effect of "fanciness," such that fancy lawyers had significantly higher comprehension overall than non-fancy lawyers (β =.433,SE=.158,p=.006). There were no other main effects in the two models.

C.2. In-Group Signaling. With regard to the in-group signaling hypothesis, We added the demographic factors to our model of hireability, as follows:

- `clmm( hireability ~ condition + age + is fancy + gender + ethnicity + practice experience + (1 + condition | item) + (1 + condition | subject), data = ., family = binomial(link = "logit"))`

As with our pre-registered model, this model revealed a main effect of register, such that lawyers rated authors of simple contracts as significantly more hireable than authors of legalese contracts (β =1.900,SE=.322,p<.0001).

C.3. It's Just Business. With regard to the it's just business hypothesis, We added the demographic factors to our model of willingness to sign, as follows:

- `clmm( client would sign ~ condition + age + is fancy + gender + ethnicity + practice experience + (1 + condition | item) + (1 + condition | subject), data = ., family = binomial(link = "logit"))`

As with our pre-registered model, this model revealed a main effect of register, such that lawyers predicted that clients would be significantly more likely to sign simple contracts than legalese contracts (β =1.208,SE=.370,p=.001).

C.4. Complexity of Information. With regard to the complexity of information hypothesis, We added the demographic factors to our model of enforceability, as follows:

- `clmm( enforceability ~ condition + age + is fancy + gender + ethnicity + practice experience + (1 + condition | item) + (1 + condition | subject), data = ., family = binomial(link = "logit"))`

As with our pre-registered model, this model revealed no effect of register. That is, we did not find evidence that lawyers rated legalese contracts as more enforceable than legalese contracts (p=.156).

C.5. Copy and Paste. We added the demographic factors to our models of quality and usability, as follows:

- `clmm(quality ~ condition + age + is fancy + gender + ethnicity + practice experience + (1 + condition | item) + (1 + condition | subject), data = ., family = binomial(link = "logit"))`
- `clmm(would use ~ condition + age + is fancy + gender + ethnicity + practice experience + (1 + condition | item) + (1 + condition | subject), data = ., family = binomial(link = "logit"))`

In both cases, we still found a main effect of register on responses, such that lawyers were significantly more likely to say that they would use simple contracts over legalese contracts (β =1.533,SE=.285,p<.0001), and rated simple contracts as significantly higher quality than legalese contracts (β =1.807,SE=.338,p<.0001).

Table S1. Endorsement rates by desiderata

Desiderata	Legalese			Simple		
	endorsement %	lower CI	upper CI	endorsement %	lower CI	upper CI
hireability	31.4	28.8	33.9	59.4	56.7	62.1
likelihood of being signed	69.0	66.4	71.4	82.2	80.2	84.4
enforceability	82.3	80.2	84.4	84.3	82.2	86.2
quality	2.61	2.55	2.67	3.33	3.27	3.39

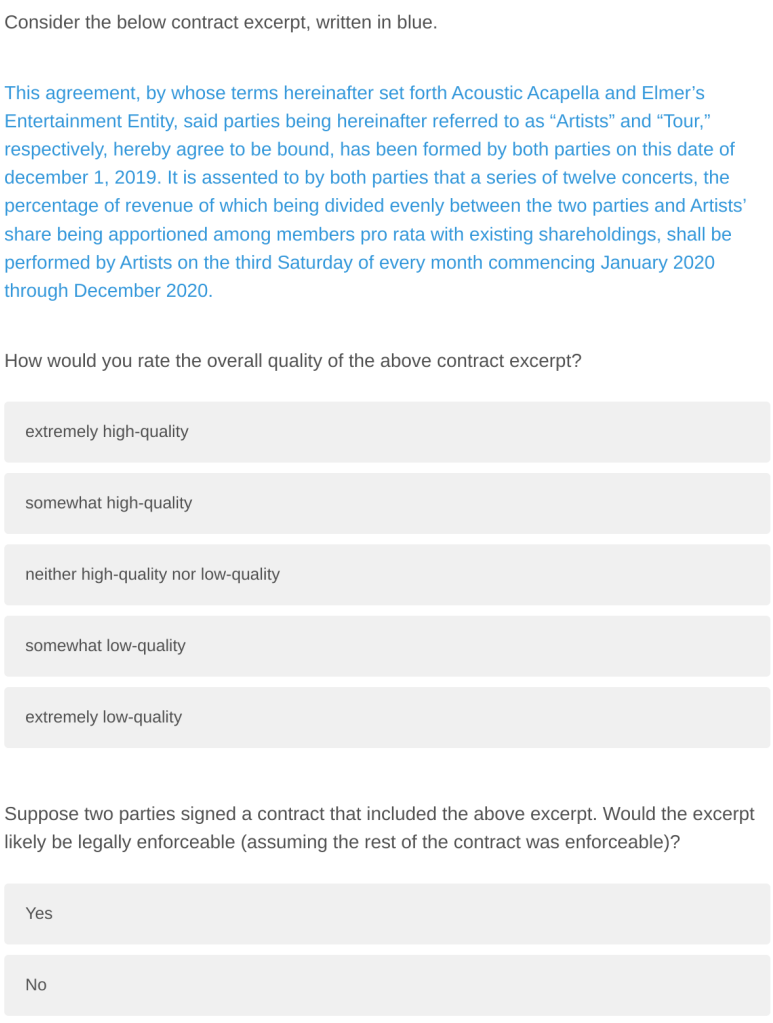

Fig. S1. Interface of Experiment II.

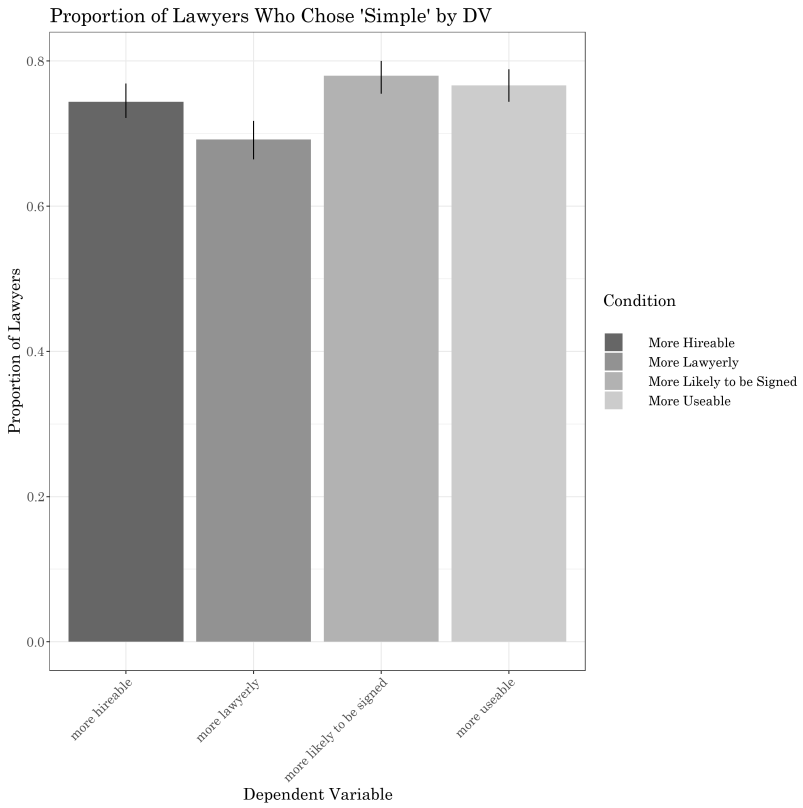

Fig. S2. Proportion of lawyers who endorsed simple version over legalese version according to different desiderata.

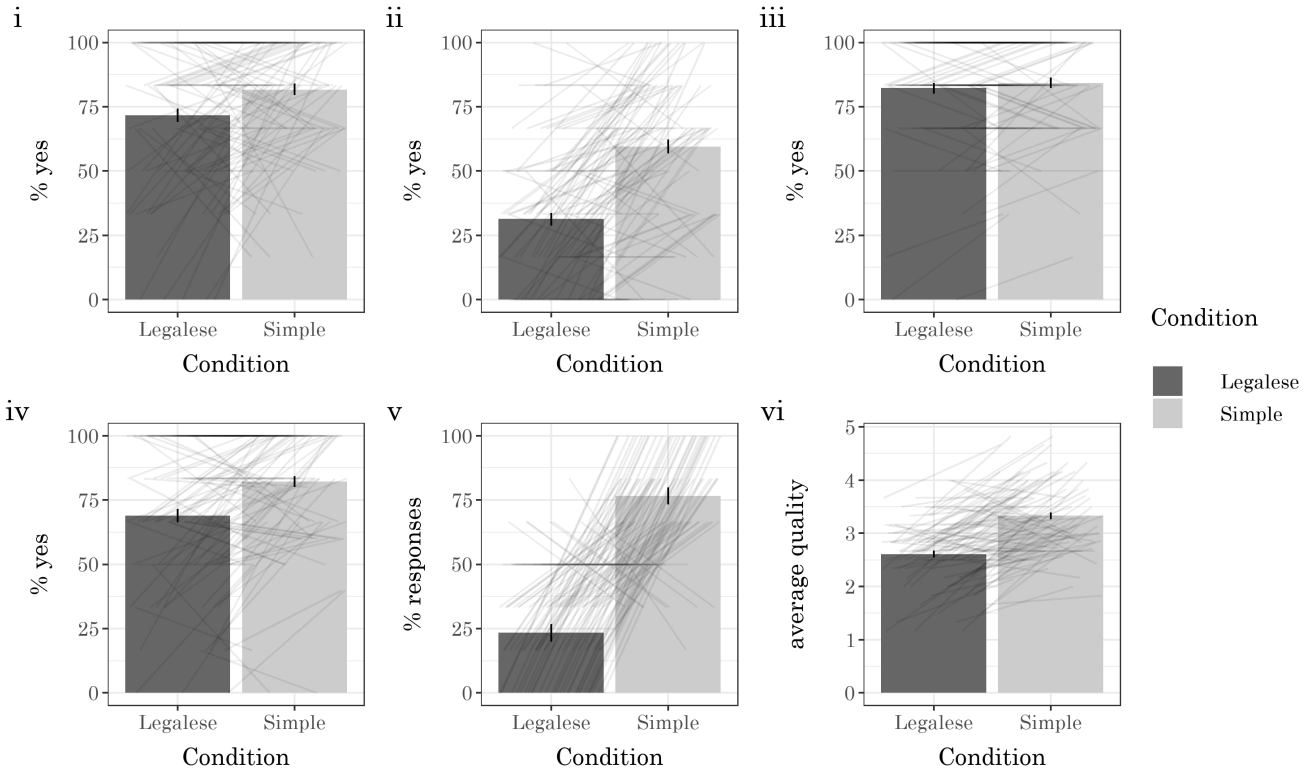

Fig. S3. Proportion of lawyers who endorsed simple and legalese contracts according to different desiderata.

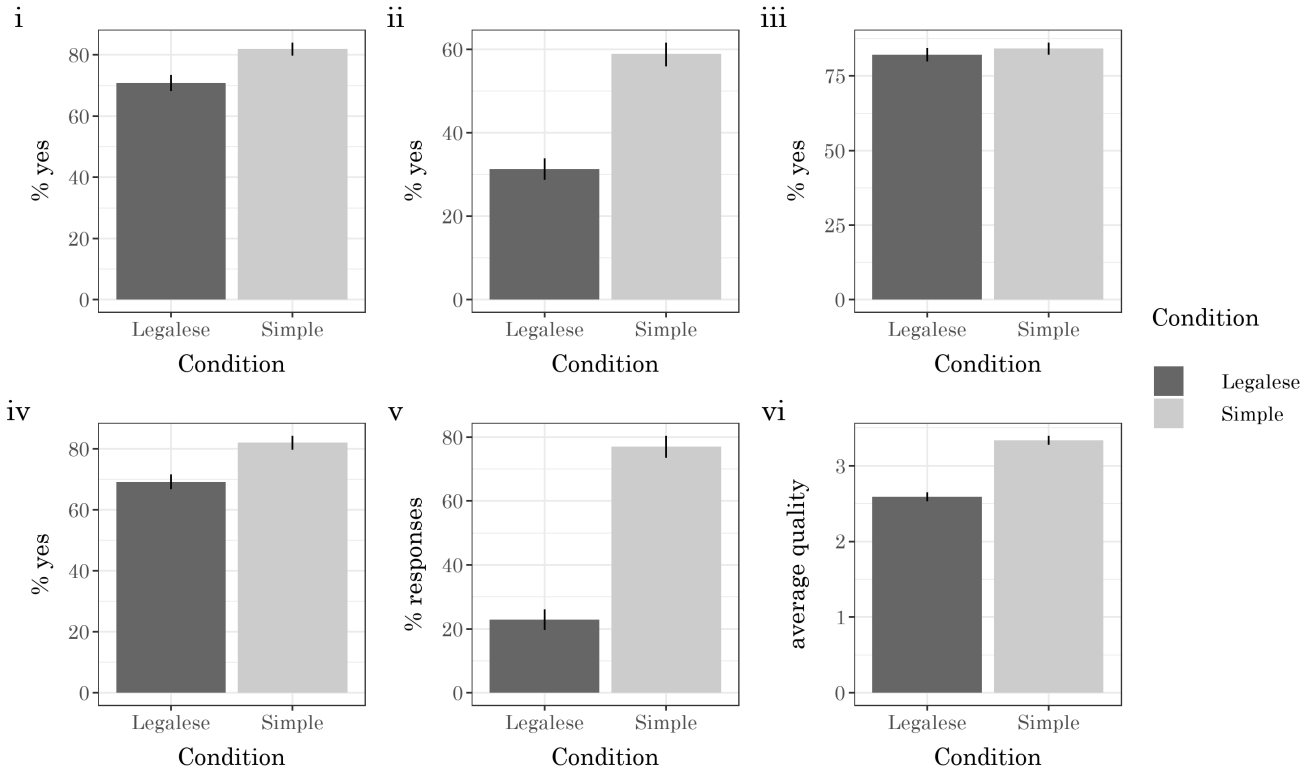

Fig. S4. Proportion of experienced lawyers who endorsed simple and legalese contracts according to different desiderata.

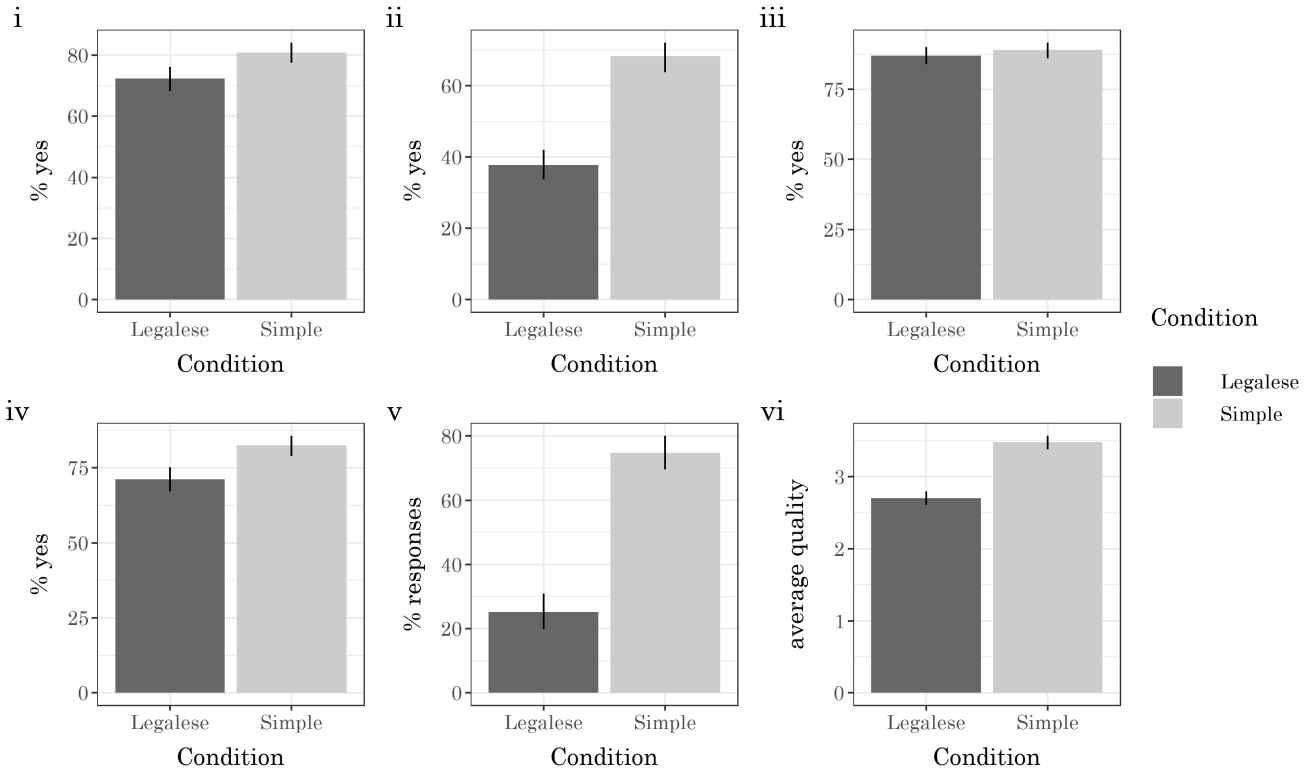

Fig. S5. Proportion of fancy lawyers who endorsed simple and legalese contracts according to different desiderata.

References

1. E Martínez, F Mollica, E Gibson, Poor writing, not specialized concepts, drives processing difficulty in legal language. *Cognition* **224**, 105070 (2022).
2. J Cohen, A coefficient of agreement for nominal scales. *Educ. psychological measurement* **20**, 37–46 (1960).
3. ML McHugh, Interrater reliability: the kappa statistic. *Biochem. medica* **22**, 276–282 (2012).
4. P Tiersma, Some myths about legal language. *J. Law, Cult. Humanit. Forthcoming, Loyola-LA Leg. Stud. Pap.* (2005).
5. P Tiersma, The nature of legal language in *AILA applied linguistics series: Vol. 5. Dimensions of forensic linguistics.* (John Benjamins Publishing Company), pp. 7–25 (2008).
6. CA Hill, A comment on language and norms in conplex business contracting. *Chi.-Kent L. Rev.* **77**, 29 (2001).
7. D Mellinkoff, *The language of the law.* (Wipf and Stock Publishers), (2004).