Using stepwise algorithm to select features.

---

**PARTICIPANTS**
- Native English speakers (4th-6th grade, 16 boys) from Falk Laboratory [56].

**TASK**
- Students provided Familiarity Ratings for 60 SAT word levels (ratings: 53%, 23%, and 26%).
- Familiarity Ratings were followed by the Meaning-Generation task (Fig 2).

**OUTCOME MEASURE**
- Two human raters labeled “off-task” responses (inter-rater agreement was Kappa = 0.695).
- Instructions (based on Baker et al. [1]): “The response seems less serious or less relevant for a given target word.”
- “The response was part of repetitive responses over different question items.”
- “The response was part of repetitive false submissions”.

**FEATURES**
- Real-time variables (RTVs): Features obtained from a single response.
- Context-based variables (CTVs): Features from historical responses and other students.
- Using stepwise algorithm to select features.

---

**ANALYSIS & RESULTS**

### FIXED EFFECT RESULTS

- Real-time variables (RTVs):
  - RT_Making: Shorter response time for typing in more likely to be an off-task behavior.
  - RespLength: Shorter responses were more likely to be an off-task behavior.
  - OrthoOverlap: Responses that were orthographically similar to the word were more likely to be labeled as an off-task behavior.
  - SpellErr: # of spelling errors.
  - RT_Start: Time spent before initiating the response.

- Context-based variables (CTVs):
  - SemanticDistance_prev.3: Responses that were semantically similar with previous 3 responses were less likely to be an off-task behavior.
  - OrthoRepetition_prev.7: Responses that were orthographically similar with previous 7 responses were less likely to be an off-task behavior.
  - pFlag_prev.X: Proportion of off-task responses in previous task.
  - TargetFlags_prev.X: Average proportion of off-task responses for previous X trials from other students.
  - TargetFlags: Proportion of off-task words from other students.

---

### RANDOM EFFECT RESULTS

- Familiarity: Words rated as unknown were more likely to elicit off-task responses.
- Variability across items & students: highlights the importance of models that capture multiple sources of variance, including random as well as fixed effects.

**REFERENCES**


---

**ACKNOWLEDGMENTS**

The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, Grant R324A160067 (to) J.P. Collins-Thompson & Frishkoff. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. We thank Falk Laboratory School for frank collaboration. We also thank Russell Brand, Mark H. Heron, Janiener Astrid, Austin Murcns and Guda Tenjons for assistance with the tutor and treadmill development.

---

**FUTURE WORK**

- Develop an adaptive vocabulary learning system
- Adaptive system that can minimize the off-task behaviors during the learning task.
- Find desirable difficulty level for each student.
- Identify behavioral log features related with perceived difficulty.

**LABELING FROM NON-EXPERTS**

- Fragmentary job for anonymous workers.
- Require more careful design instructions.

---

**CONTRIBUTION**

- Methods for extracting meaningful information from log data.
- RTV + CTV with mixed effect model.
- CTVs can substitute traditional off-task predictive features, such as # of error messages and response time.

**IDENTIFYING OFF-TASK STATUS AT THE ITEM LEVEL**

- Letting the learning system know when to intervene.
- Manage student engagement systematically.

**MORE ACCURATE PREDICTION ON THE STUDENT’S VOCABULARY KNOWLEDGE**

- Distinguishing between accidentally erroneous responses and intentionally missed responses.