Human Understanding of Controlled Natural Language in Simulated Tactical Environments

Erin Zaroukian
Human Research and Engineering Directorate
US Army Research Laboratory
Aberdeen Proving Ground, USA
erin.g.zaroukian.ctr@mail.mil

***Abstract*—Computational platforms with natural language interfaces have become commonplace, but they present limitations that make them less than ideal for military and other safety-critical environments. Controlled Natural Languages (languages built from a subset of natural language, which are both computer- and human-readable) hold promise for Human-Computer Collaboration via these platforms, especially when the human user needs to add information to a knowledgebase or make queries, as they provide a transparent, shared representation. Controlled Natural Languages, however, are typically not optimized for human use and understanding. This paper presents the development and implementation of a framework to test the relative ease of comprehension of different Controlled Natural Language statements. The experiments presented in this paper show an advantage for one particular Controlled Natural Language statement over another, but only when responses are made under strict time pressure. These types of experiments allow researchers to make recommendations on how to improve the use and design of a Controlled Natural Language for more robust comprehension, particularly in tactical environments.**

Keywords—human-computer collaboration; Controlled Natural Language

I. Introduction

Human-computer collaboration (HCC), where humans and computer agents work together to achieve a shared goal, has great potential to increase situational awareness and improve performance on crucial military tasks. In the Intelligence Cycle [1], for example, computers can aid in accessing and recording information, and even analyzing collected information and making decisions. Technology using a simple conversational interface between the human user and the computer agent can be used to aid the allocation of intelligence, surveillance, and reconnaissance (ISR) resources [2], as well as to aid data collection and build a knowledge base [3]-[4]. While users in these studies can enter information and queries using natural language, much of the collaboration takes place using a Controlled Natural Language (CNL), which is both human- and computer-readable. A CNL acts as a shared representation between the human user and the computer agent. This avoids the error-prone task of parsing NL into a computer-readable form, and it allows the computer to answer queries and perform computational reasoning over its knowledge base, responding with transparent rationale.

There are clear advantages to optimizing CNLs for human understanding, but the empirical research required to inform such work is lacking. This paper takes a step toward improving said research. After an overview of CNL and relevant research to-date, this paper describes 2 experiments examining human comprehension of a CNL, providing direct comparisons in accuracy and response time among CNL statements.

II. Controlled Natural Language

A CNL is "… a subset of natural language that can be accurately and efficiently processed by a computer, but is expressive enough to allow natural usage by non-specialists" [5]. Because a CNL provides a shared representation between the human user and the computer agent, both the information used by the system and the rationale for decision making are transparent to the human user, obviating black-box algorithms and fostering trust.

Results from a Simple Human Experiment Regarding Locally Observed Collective Knowledge [3] show untrained users productively communicating with a computer agent via CNL to build a knowledge base, suggesting that CNLs work well as a shared representation. Little is known, however, about just how easily human users understand a given CNL.

Large strides toward assessing human comprehension of CNLs were made by Kuhn [6]-[7], who developed a framework for evaluating and comparing comprehension of CNLs. In this framework, participants were asked to judge a CNL statement as true/false. The truth of the statement was determined relative to a provided ontograph, which is a graphical notation Kuhn developed for representing ground truth. Ontographs represent a closed world, where shown entities and relations are the only ones that exist (e.g., in Fig. 1 there is no person named Paul, Bill is not an officer, and Lisa does not see Tom). In one study in Kuhn's framework [6], participants showed overall high accuracy (~85% correct) in their responses to CNL statements. Another study [6] used a similar true/false task to compare participants' comprehension of a CNL to their comprehension of a simplified ontology, and participants in this study showed higher accuracy with the CNL statements.

Dr. Zaroukian was supported by an appointment to the US Army Research Laboratory Postdoctoral Fellowship Program administered by the Oak Ridge Associated Universities.

While previous studies provide a starting point, they examine only accuracy, not speed, of comprehension. Speed indicates ease of comprehension, with faster response times indicating quicker and less effortful processing for comprehension. This is important in cases where participants may be performing with high accuracy across items because the response times for different items may vary systematically, indicating that the items are not all equal in terms of ease of comprehension. Additionally, previous studies have not attempted to inform the design or use of CNL for improved human comprehension.

While CNLs have great potential in HCC, behavioral research measuring speed and accuracy is required. Such research will help determine how efficient CNLs are to human users, and will help develop principles for CNL design and use. To this end, the following experiments develop a framework, drawing on Kuhn's ontograph approach, for testing paraphrases within ITA Controlled English (CE) [9]. CE is a CNL developed as part of the International Technology Alliance and used in [2]-[4]. In CE, it is simple to define entities and rules, allowing for many potential paraphrases that express the same information. For example, if the domain model contains the relation "sees" such that a user can state "the person John sees the person Tom," the user can add "is seen by" to the model in order to convey the same information as, "the person Tom is seen by the person John." In the following experiment, 3 paraphrases were tested to determine if certain phrasing was more quickly and accurately comprehended within the given context.

III. Experiment 1[1]

Experiment 1 presents an initial study using a framework similar to Kuhn's, where participants judged a statement as true/false relative to a provided diagram. Unlike Kuhn's studies, however, this experiment directly compared different ways of expressing the same information within the same CNL, and it measures both speed and accuracy as indicators of comprehension. The aim is to use participants' responses to these different paraphrases to determine whether certain expressions are more quickly or accurately understood and should be recommended for use above others. In this experiment, the comparison is among 3 ways of asserting that 2 entities are unique: roughly, "X is not Y", "X cannot be Y", and "X is different to Y".

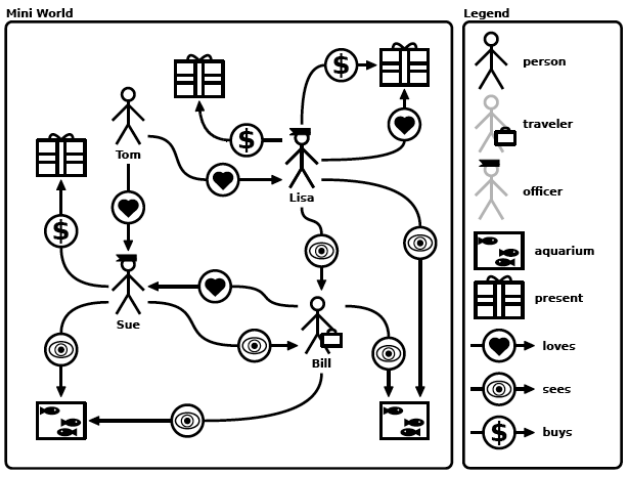

Fig. 1. Example of an ontograph from [6].

[1] This work was presented before data collection was complete at the Annual Fall Meeting of the International Technology Alliance in College Park, MD [10].

A. Participants

Seventy-five participants were recruited through Amazon Mechanical Turk and were paid $0.75 for their participation. Mechanical Turk directed participants to Ibex Farm [11]-[12], which hosted the experiment. A demographic survey was given before the experiment began, showing that participants were between the ages of 21-64, 29 were female, 65 were native English speakers, and they had varying experience with logic/programming: No knowledge, n=33; A Little Knowledge, n=18; Some Knowledge, n=17; A Lot of Knowledge, n=3; and Expert knowledge, n = 4.

B. Materials and procedures

After the survey, participants were presented with a rule written in CE paired with a diagram, and for each pair they were asked to respond (Yes/No) to the question "Is the diagram consistent with the rule?" An example is shown in Fig. 2. Participants worked through 7 labeled training items, followed by 24 test items.

a) Diagrams

The diagrams in this study were modeled after Kuhn's ontographs and represent closed worlds. All diagrams contained 3 people (John, Mary, Peter), 3 books (War and Peace, Middlemarch, Moby-Dick), and reading relations represented by arrows. Four diagrams with relatively simple relations were used in practice, and 4 diagrams with more complex relations were used in test.

b) Rules

Uniqueness, the contrast of interest in this experiment, was expressed in 3 ways, exemplified below:

"the person John is not the person Tom"

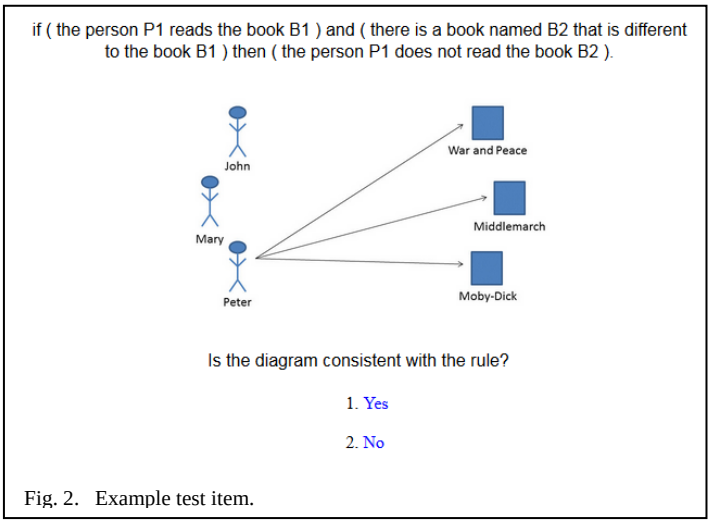

Fig. 2. Example test item.

"the person John cannot be the person Tom"

"there is a person named John that is different to the person Tom"

All rules were of the form "if (STATEMENT) and (STATEMENT) then (STATEMENT)". Some of the STATEMENTs were like the examples above, but contained variables (such as P1, P2) instead of specific names (see Examples 1-4, below).

To create variety in the items and discourage participants from developing superficial response strategies, rules were varied in a number of ways. Examples of the 4 types of uniqueness items are provided in (1-4) below using "is not" (for clarity, natural language translations are given below each example here, but these were not provided to participants).

1. if (the person P1 reads the book B1) and (the book B2 is not the book B1) then (the person P1 does not read the book B2).

   'If a person reads a book, that person does not read any other book.'

2. if (the person P1 reads the book B1) and (the book B2 is not the book B1) then (the person P1 reads the book B2).

   'If a person reads a book, that person reads every other book too.'

3. if (the person P1 reads the book B1) and (the person P2 is not the person P1) then (the person P2 does not read the book B1).

   'If a person reads a book, no other person reads that book.'

4. if (the person P1 reads the book B1) and (the person P2 is not the person P1) then (the person P2 reads the book B1).

   'If a person reads a book, every other person reads that book too.'

Participants also saw 2 types of fillers, shown in (5-6).

5. if (the person P1 reads the book B1) and (the person P2 reads the book B2) then (the book B2 is the book B1).

   'If someone reads a book and someone (possibly the same person) reads a book, then those books are the same book.'

6. if (the person P1 reads the book B1) and (the person P2 is the person P1) then (the person P2 reads the book B2).

   'If a person reads a book, and there is another person who is actually the same person, then that person reads a book (possibly the same as the first book).'

Furthermore, the order of the 2 statements in the "if"-clause were shown in the order above, as well as reversed.

c) Procedure

Each participant began with a survey, then 7 practice items. These practice items contained simplified diagrams and rules intended to introduce the statement–diagram paradigm, and teach them how to read CE rules and interpret CE variables. The practice items did not contain any uniqueness expressions. Participants were given step-by-step instructions on how to solve 4 of the practice items. Upon responding to any practice item, participants were told whether their response was correct or incorrect and were given an explanation of how to solve that item.

After completing the practice items, participants saw 24 test items, each separated by a rest screen reminding them to respond as quickly and as accurately as possible. Participants were not told whether their answer was correct. Sixteen of these 24 items contained the contrast of interest, while 8 were fillers, which were not included in analyses. Each participant saw each of the 4 main rule types (1-4) with each of the 4 diagrams: 2 that made it true, and 2 that made it false. A Latin square design determined which uniqueness expression was used in each rule, and the number of regular/reversed antecedents was balanced within subjects.

C. Results

Mean accuracy was 0.747 (SE=0.031) with mean reaction time, 12.802s (SE=1.309s), represented for individuals in Fig. 3. The contrast of interest between the uniqueness expressions is shown in Fig. 4. A generalized linear mixed model[2] [14]-[16] with worker as random effect and rule form as fixed effect[3] revealed no effect of uniqueness expression on accuracy ($\chi 2(2)=0.727$, $p=0.695$) or response time ($\chi 2(2)=0.848$, $p=0.654$).

D. Discussion

High mean accuracy suggests that this experimental framework allows participants to demonstrate comprehension of CE. While a number of participants' accuracy is at ceiling, a similar number are at chance (0.5), indicating that, despite high overall accuracy, some participants do not understand the task fully. Absolute accuracy, however, is not of primary interest here; relative accuracy among different uniqueness expressions

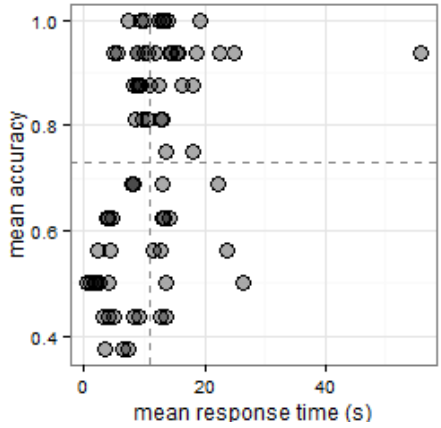

Fig. 3. Mean accuracy and response time per worker in Experiment 1, with overall means as dashed lines.

[2] Generalized mixed models are a generalization of linear regression that can fit non-normal dependent measures by including both fixed and random effects [13].

[3] While the addition of rule form as fixed effect did increase the plausibility of the model, rule form was nonetheless part of the experimental manipulation and is included in the models reported [17].

is of much greater interest. Since *why* certain participants performed at chance cannot be determined (e.g., difficulty understanding the diagrams, understanding the rules, using their own computer), differences in (lack of) comprehension of uniqueness expressions are difficult to interpret. Thus, Experiment 2 focused on data from participants who performed above chance.

Response times provide 2 insights. First, a number of participants had very short response times, which tended to coincide with low accuracy, suggesting that these participants did not carefully read the CE rules and/or inspect the diagram. Thus, in Experiment 2, only data from participants with above chance accuracy—i.e. who likely understood the task—are analyzed. Second, inspection of individual trials reveals a number of long response times, well beyond 2 standard deviations above the mean (26.75s). Because this experiment showed high overall accuracy, with many participants at ceiling, Experiment 2 introduced a time limit, which may lower accuracy and may be crucial for identifying performance differences among the uniqueness expressions.

IV. Experiment 2

Experiment 2 was identical to Experiment 1, with the addition of a time constraint. It was hypothesized that a time constraint would lower accuracy and reveal differences in performance within the contrast of interest—uniqueness.

A. Participants

A time limit is likely to lower performance, and it was unclear how many participants would be required to retain adequate power after removing those with mean accuracy at 0.5 and below. To inform this, an initial 77 participants[4] were recruited via Mechanical Turk to pilot Experiment 2. These participants showed a mean accuracy of 0.544 (SE=0.020) and a mean response time of 6.189 (SE=.0411). Of these participants, 37 had a mean accuracy above 0.5.

Informed by this preliminary study, an additional 199 participants were recruited through Amazon Mechanical Turk for Experiment 2. The demographic survey showed that participants were between the ages of 18-74 years, 105 were female, 168 were native English speakers, and they again had varying experience with logic/programming: No knowledge, n=92; A Little Knowledge, n=32; Some Knowledge, n=49; A Lot of Knowledge, n=18; and Expert knowledge, n = 8.

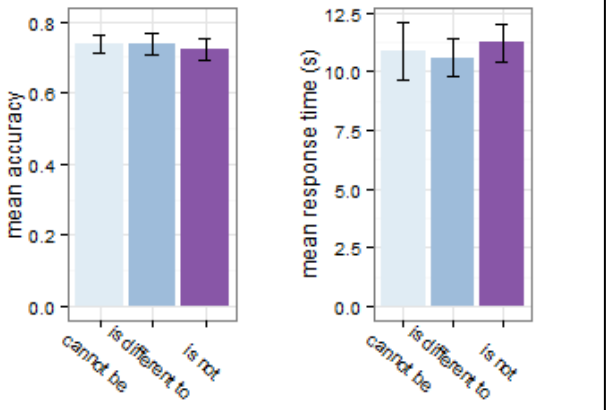

Fig. 4. Mean accuracy and response time by uniqueness expression in Experiment 1, with SE.

[4] While 75 participants were targeted for this preliminary study and 200 for Experiment 2, idiosyncrasies in the communication between Mechanical Turk and the server hosting the experiment led to slight discrepancies.

B. Materials and procedure

Materials and procedure in Experiment 2 were the same as Experiment 1, but with an added time limit. Participants were told that if they did not complete an item within 15s, the task would automatically progress to the next item. They were reminded of this time limit in the rest screen between items. If participants did not respond within the time limit, the trial was considered incorrect.

C. Results

For the 132 participants that scored above chance (0.5), mean accuracy was 0.713 (SE=0.011) and mean response time 8.106s (SE=0.264s). Data for all 199 participants is shown in Fig. 5. The contrast of interest between uniqueness expressions is shown in Fig 6. A generalized linear mixed model with worker as random effect and rule form as fixed effect[5] revealed a significant effect of uniqueness expression on accuracy ($\chi 2(2) = 6.485$, p=0.039), but no significant effect on response time ($\chi 2(2)=3.695$, p=0.157). This gives an evidence ratio of 3.36. Between-group comparisons based on least squares means show that only "cannot be" and "is not" are significantly different, p=0.032 [18].

D. Discussion

While Experiment 1 failed to show differences in ease of comprehension among the 3 paraphrases tested, it revealed wide ranges both in response times and in accuracy across participants. Long response times might mask difference in performance, and chance performance suggests failure to understand the task. Experiment 2 addressed these issues by introducing a time limit and by restricting analysis to those

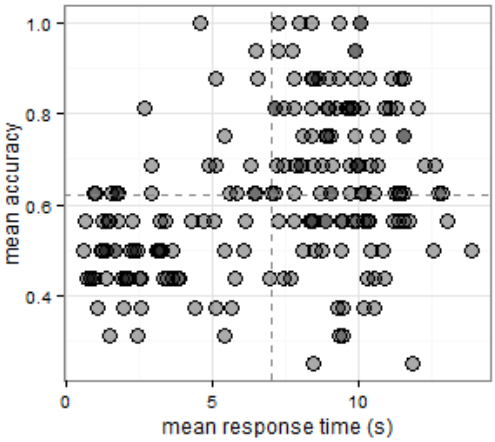

Fig. 5. Mean accuracy and response time per participant in Experiment 2, with overall means as dashed lines.

[5] Again, while not increasing model plausibility, rule form is included.

who performed above chance (i.e., likely understood the task). This revealed a small difference in participants' accuracy on the different uniqueness expressions, where responses to "is not" were more accurate than responses to "cannot be".

On one hand, this result is unsurprising, as the verb "cannot" is often used in statements made via inference—e.g., "John isn't home (I know this for a fact)" vs. "John can't be home (because I just saw him at the store)". If participants are trying to accommodate this interpretation by inventing a context that fits with an inferential reading, this could lead to slower response times or lower accuracy. Participants may even find such rules uninterpretable in this context. On the other hand, these results are surprising. The word "not" is crucial but easy to miss, especially given that the experiment contained filler items with bare "is" (not with bare "can"). Negation is also easy to misremember [19], but no advantage was found for the more salient, positive "is different to".

V. Conclusion

In tactical environments, where decision making is done with limited time resources, CNLs hold great potential. They serve as a shared representation between human and computer agent, sidestepping the unreliable automated step of translating NL to a computer-readable format while providing support for improved situational awareness. Further, CNLs can be directly shaped to the user's needs and limitations. The experiments presented in this paper show that, in a situation with strict time constraints, the precise phrasing of a statement can make a significant difference in the accuracy of the human user's comprehension of that statement. This was seen in Experiment 2, which introduced a time constraint and showed differences in accuracy where "is not" produced more accurate responses than "cannot be". This moves a step toward making evidence-based recommendations on how CNLs should be designed and used in these contexts.

Acknowledgment

This research was supported in part by an appointment to the U.S. Army Research Laboratory Postdoctoral Fellowship Program administered by the Oak Ridge Associated Universities through a cooperative agreement with the U.S. Army Research Laboratory. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. Special thanks are due to Jon Bakdash, David Mott, Alun Preece, Dave Braines, and Katherine Gamble.

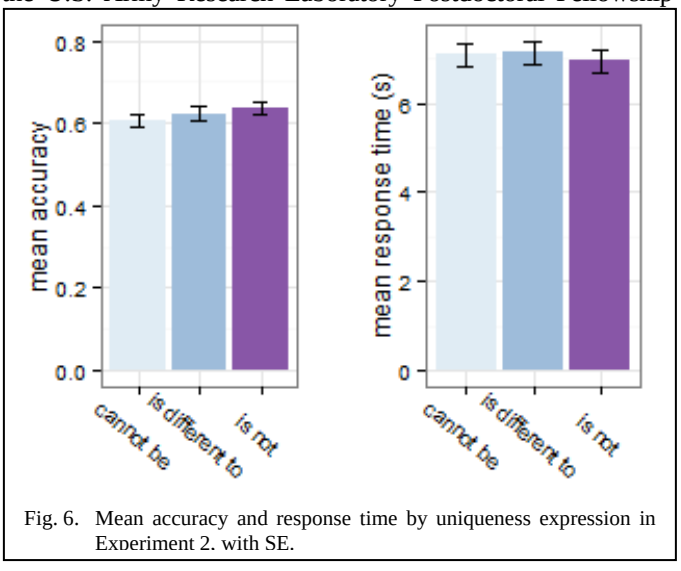

Fig. 6. Mean accuracy and response time by uniqueness expression in Experiment 2, with SE.

References

[1] "Army Doctrine Reference Publication No. 2-0 (FM 2-0)," Headquarters, Department of the Army, Aug. 2012. http://armypubs.army.mil/doctrine/DR_pubs/dr_a/pdf/adrp2_0.pdf

[2] D. Pizzocaro, C. Parizas, A. Preece, D. Braines, D. Mott, and J. Z. Bakdash, "CE-SAM: A conversational interface for ISR mission support," in *SPIE Defense, Security, and Sensing*, Baltimore, MD, 2013.

[3] A. Preece, W. Webberley, D. Braines, N. Hu, T. La Porta, E. Zaroukian and J. Z. Bakdash, "SHERLOCK," presented at *Annu. Fall Meeting of the Internat. Technology Alliance*, College Park, MD, 2015.

[4] A. Preece, C. Gwilliams, C. Parizas, D. Pizzocaro, J. Z. Bakdash, D. Braines, "Converational Sensing," in *Proc. Next-Generation Analyst II (SPIE Vol 9122)*, SPIE, 2014.

[5] N. E. Fuchs, R. Schwitter, "Specifying Logic Programs in Controlled Natural Language," presented at *Workshop on Computational Logic for Natural Language Process.*, Edinburgh, UK, 1995.

[6] T. Kuhn, "How to evaluate controlled natural languages," in *Pre-Proc Workshop on Controlled Natural Language*, Marettimo Island, Italy, 2009.

[7] T. Kuhn, "Controlled English for Knowledge Representation," Ph.D. dissertation, Faculty of Econ., Bus. Admin. and Inform. Technology, Univ. Zurich, Zurich, Switzerland, 2010.

[8] T. Kuhn, "The understandability of OWL statements in controlled English," *Semantic Web*, vol. 4, no. 1, 2013.

[9] D. Mott, "Summary of Controlled English," unpublished.

[10] D. Mott and E. Zaroukian, "Studies in the Human Use of Controlled English," presented at *Annu. Fall Meeting of the Internat. Technology Alliance.*

[11] A. Drummond, *Ibex* (version 0.3), https://code.google.com/p/webspr/, 2013.

[12] A. Drummond, *Ibex Farm*, http://spellout.net/ibexfarm.

[13] C. E. McCulloch and S. R. Searle, "Linear Mixed Models," in *Generalized, Linear, and Mixed Models*. Hoboken, NJ: Wiley, 2001, ch. 6, pp. 156-186.

[14] B. M. Bolker, M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J. S. S. White. "Generalized linear mixed models," in *Trends in ecology & evolution*, vol. 24, no. 3, pp. 127-135, 2009.

[15] D. Bates, M. Maechler, B. Bolker, and S. Walker, *lme4: Linear mixed-effects models using Eigen and S4*, (R package version 1.1-7), http://CRAN.R-project.org/package=lme4, 2014.

[16] R Core Team, *R: A language and environment for statistical computing*, R Found. for Statistical Computing, Vienna, Austria, http://www.R-project.org/, 2014.

[17] K. P. Burnham and D. R. Anderson, *Model selection and multimodel inference: a practical information-theoretic approach*, New York, NY: Springer, 2002.

[18] R. Lenth, *lsmeans: Least-Squares Means*, (R package version 2.20-23), http://CRAN.R-project.org/package=lsmeans, 2015.

[19] R. Mayo, Y. Schul, and E. Burnstein, "'I am not guilty' vs. 'I am innocent'," *J. Experimental Social Psychology*, vol. 40, pp. 433-449, 2004.