The Importance of Random Assortment and Blinding in Qualitative Data Analysis

Abstract

Qualitative data analysis contains some degree of error, and any research group that performs qualitative research should be aware of sources of bias. The Automated Analysis of Constructed Response (AACR) research group investigates computerized analysis of undergraduate students' constructed responses in science and statistics. In our development of computer models, increasing the number of human-coded responses enhances the accuracy of these models but analysis techniques that decrease coding time, such as confirming computer-predicted codes instead of coding without that information, can inject new sources of bias. We tested whether having the computer-predicted codes visible (i.e., no blinding) and having the responses in the order of the computer codes (i.e., no random assortment) created bias in human coding. In the initial investigation, one coder coded 2,000 responses to a three-part constructed response question and found an effect from both no blinding and no random assortment. To test whether this finding was a novel observation, three coders including the initial coder replicated the study with experimental improvements. Coders were aware of the overall pattern of the initial coder's responses. Contrary to the previous findings, random assortment and blinding had little to no effect on bias. Our results suggest that blinding and random assortment may be effective methods for reducing bias in coding student responses but may be less important when other methodological aspects are rigorous.

Introduction

No matter the research question, data analysis contains some degree of error. If we examine a distribution of data values, the ideal situation is that the error is randomly distributed so that they are equally represented about the true mean. In this situation, even though error occurs, error on one side of the true mean cancels out the other side, thereby making the results still reliable as the true mean is represented (Rosenthal 1976). Bias, on the other hand, causes error that is not equally distributed about the true mean. Instead, the error is larger in magnitude on one side than the other, which distances the sample mean from the true mean. In considering qualitative data analysis, the investigation may not be about discovering a numerical mean but still seeks the truth behind the research question, and bias can alter the estimation of that truth.

Bias can occur in both the examined participants and the researchers that are performing the examination, and for participants and researchers bias often occurs by knowing more about the methods than what is necessary to participate in the study. For instance, participants may know into which treatment group they have been placed and researchers may know the background of their participants. If researchers know more about their subjects than is required to rate them, for instance, their codes may be due to information that confirms what they already believe they know- an effect called confirmation bias, which is a type of cognitive bias.

Because some degree of error is inevitable. Biases like conformation bias, (and other cognitive biases) are avoidable if handled correctly. A possible way to reduce bias is being aware of one's own bias (Rosenthal 1976). Another process that may reduce researcher bias is performing replicates of the study (Rosenthal 1976). Rosenthal (1976), however, cautions that this process may not be enough if the same methodological error is repeated. Also, having a large

sample size may reduce error but likely reduces only random, non-systematic error (Guyatt & Furukawa 2008).

Blinding, which entails removing information about the units being analyzed such as which units are receiving which treatment, is a possible method for minimizing bias. In an automated meta-analysis of over 7,600 life science peer-reviewed articles, the use of blinding negatively correlated with statistical significance; in other words, studies that implemented blinding techniques in their methods were less likely to have significant findings than those that did not use blinding (Holman et al. 2015). Confirmation bias can also occur during teaching and grading (Hofer 2015; Malouff et al. 2013; Rosenthal and Jacobson 1968)

Another method to reduce bias is randomization. When units are in the order of a variable, such as treatment type or other subject characteristics, it can cause order effects, which occurs when units with the same variable are coded or assessed more consistently by researchers when they are grouped together and/or the variables are in a particular order rather than when all units of the study are randomly sorted. Several studies have shown order effects for a variety of cases, including taking written surveys (Chan et al. 2015; McClendon 1986; Willits and Saltiel 1995), reviewing conference proposals (Cabanac and Preuss 2013), and judging wine (Mantonakis et al. 2009), counselors (Newman and Fuqua 1992), and figure skaters (Bruine de Bruin 2005). Meanwhile, a few studies have shown no effects of ordering (e.g., Butler and Cantrell 1986).

The studies described above, using quantitative or qualitative methods, illustrate several ways that bias may artificially increase classification error. Any research group should be concerned about reducing bias, and the Automated Analysis of Constructed Response (AACR) Research Group is no exception. AACR investigates computerized analysis of undergraduate

students' constructed responses pertaining to science and statistics in order to provide faculty with formative feedback to identify students' thinking (Urban-Lurain et al. 2015).

Whenever we create a new computerized scoring model, humans have to first code several hundred responses. These scored responses are then used to build a scoring model. The more human-coded responses used to create and revise the model, the more accurate it will likely become. Once a model is initially built, it is tested with new responses and the computer codes are compared to human coding. Reliability of the computer model is tested using inter-coder reliability- a test often used to compare human coders. If accuracy is poor, then additional human-coded responses are used to improve the computer prediction. Therefore, initially building a model and then adding more human-coded responses is a potential way to create a more reliable model, but human coding may take a considerably long time. On the other hand, having responses that have been scored by an imperfect computer model and then verified by a human may take less time. But, would having the computer score available result in bias in subsequent human coding? In other words, would coders be more unintentionally inclined to code responses (approximately) the same way as the computer codes? We addressed this question by testing whether having the computer codes visible (i.e., no blinding) and having the responses in the order of computer codes (i.e., no random assortment) would create bias in human coding. Since both blinding and random assortment are common methods for reducing bias, this study can inform not only our research group but the wider education research community about the effectiveness of these methods.

Methods

Data Collection

In AACR, we create computerized scoring models for constructed response questions that were developed by modifying questions primarily from concept inventories (Urban-Lurain et al. 2015). Other group members created the question used for the present analysis (see Figure 1) by modifying two existing questions from the Genetics Concept Assessment (Smith, Wood, and Knight 2008). The AACR question is composed of three parts (each part is hereafter referred to as "replication," "transcription," and "translation", respectively). We analyzed responses to each question part separately in the same order that students complete the question: replication, transcription, and then translation. The group members also created a coding scheme for the question using an inductive approach and developed a three-level holistic rubric: correct, incomplete/irrelevant, or incorrect. They created the model using 300 responses and tested the model with an additional 743 responses (Prevost et al., in preparation). In this study, we analyzed a set of over 2,000 student responses from multiple institutions per question part (over 6,000 responses total). For each response, the computerized scoring models provided the predicted code (i.e., correct, incomplete/irrelevant, or incorrect) and the probability that the code belonged in that particular category versus the other two (hereafter, simply referred to as "computer codes"). Before coding for the present analysis, Coders X and Y, who are experts in the question's subject domain, discussed and further defined the rubric and then coded 200 responses per question part (600 responses total) independently. They met the requirements of inter-coder reliability for each question part using Cohen's kappa (replication $\kappa = .80$; transcription $\kappa = .70$; translation $\kappa = .64$; according to Landis and Kock's (1977) observer scale, κ between .61 and .80 is considered substantial).

The following DNA sequence occurs near the middle of the coding region of a gene.
DNA 5' A A T G A A T G G* G A G C C T G A A G G A 3'
There is a G to A base change at the position marked with an asterisk. Consequently, a codon normally encoding an amino acid becomes a stop codon.
I. How will this alteration influence DNA replication?
II. How will this alteration influence transcription?
III. How will this alteration influence translation?

Figure 1: AACR question used in study.

In order to assess the impact of blinding and random assortment, we tested four treatments for all three question parts. These treatments are based on if the computer codes were or were not visible during human coding, and if the student responses that we were coding were in the order of the computer codes (i.e., grouped by correct, incomplete/irrelevant, and incorrect and within each code responses were in the order of the probability that the computer assigned, in descending order. The four treatments consist of:

1. Computer codes visible and responses are in the order of the computer codes (i.e., no blinding and no random assortment).
2. Computer codes are not visible and responses are in the order of the computer codes (i.e., blinding and no random assortment).
3. Computer codes visible and responses are in a random order (i.e., no blinding and random assortment).
4. Computer codes are not visible and responses are in a random order (i.e., blinding and random assortment).

We tested the four treatments in two phases.

Phase I

For each question part and each treatment, Coder X coded over 500 responses (>6,000 responses total; Coder Y created the files for coding). During this phase, we were aware of the possibility of bias but Coder X coded the responses while under the impression that the results would indicate no difference between treatments. Coder X completed treatments 1 and 4 first and then 10 weeks later completed treatments 2 and 3. The purpose for the time lapse was that the original research question was only in regards to treatments 1 and 4; after obtaining those results the research question was expanded to contain the other treatments. Coder X randomly selected the order of coding (i.e., 1 or 4 first and later 2 or 3 first) for each question part.

Phase II

Several weeks after completing Phase I, we began Phase II. Phase II was different from Phase I because it contained additional random assortment of the responses as well as additional coders: Coders X, Y, and Z (Coder Z is also an expert in the question's subject domain). All three coders were aware of the results of the previous study. Coders X and Y were trained on the rubric and established inter-coder reliability before Phase I. We trained Coder Z on the rubric using the same 200 responses for each question part. During the training, we discovered that the coding rubric was not defined enough for a coder outside of the project; therefore, we created clearer and more detailed definitions of each code during this stage. After the rubric modifications, we compared her codes to the consensus codes from Coders X and Y. Substantial agreement was found between Coder Z's codes and the consensus codes using Cohen's kappa (replication $\kappa = .75$; transcription $\kappa = .73$; translation $\kappa = .81$).

We used the same 2,000 responses used in Phase I for each question part in Phase II but this time we used a split-plot design (Figure 2), which is named as such since it was initially

developed for agricultural studies (Yates 1935). A split-plot design groups units for analysis into "plots," and then divides (or splits) each plot into subplots (Jones & Nachtsheim 2009). For our study, each treatment signifies one plot. Coder X randomly divided the 2,000 responses into the four plots (i.e., treatments). Then she randomly divided each plot into four subplots, creating 16 total subplots. She determined coding order by randomly assorting all 16 subplots. After we completed coding, we recombined the subplots in their respective plots (i.e., treatments) for data analysis. Coder X repeated this process for each question part and, while using the same student responses, for each coder.

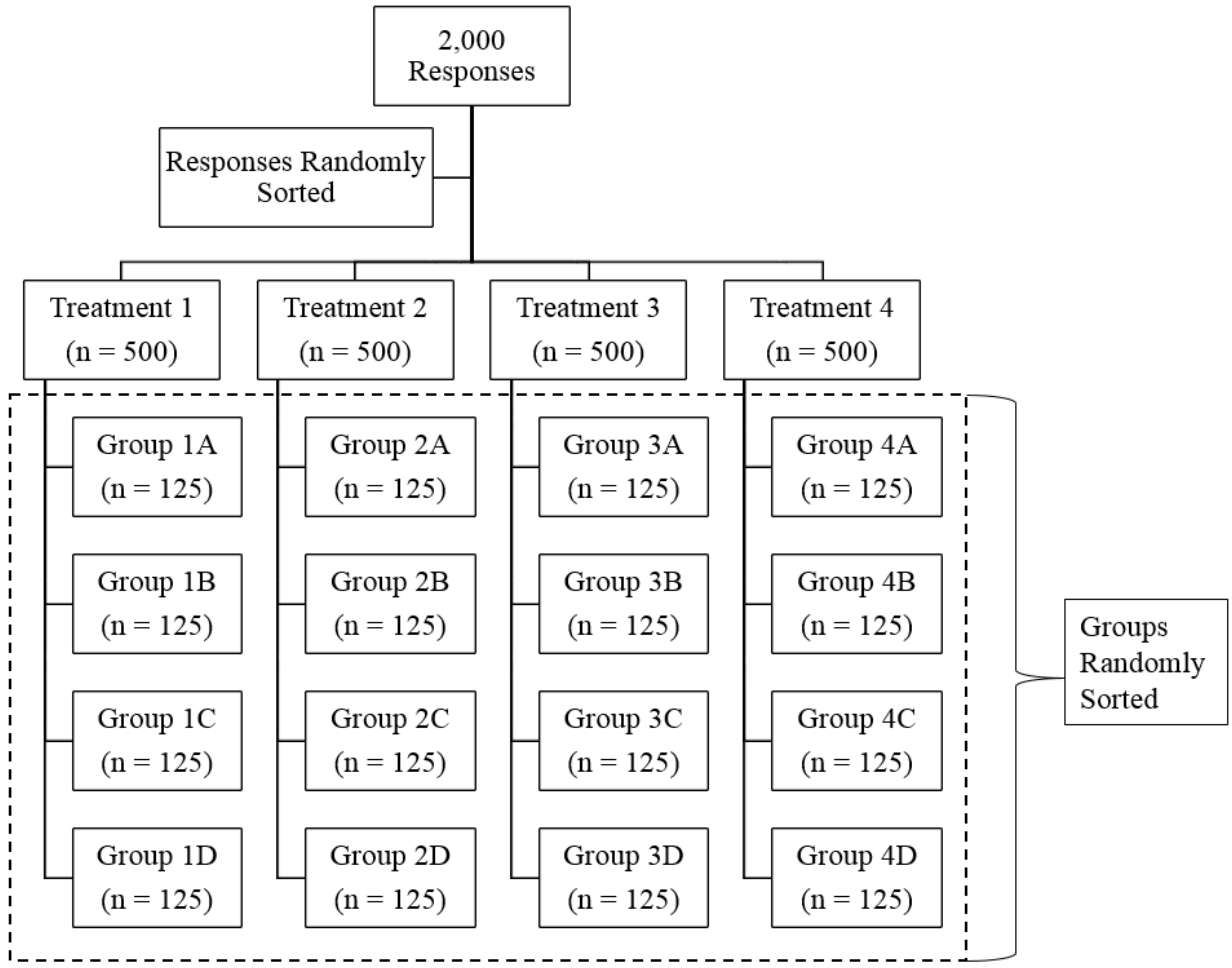

Figure 2. Split-plot design used for Phase II for each question part and each coder.

Data Analysis

For both phases and each treatment, we calculated the proportion of responses for which both the coder and the computer agreed. We expected that if either not blinding or not randomly assorting created bias, then the human codes would agree more often with the computer codes in these scenarios than when blinding or randomly assorting the responses. No significant differences in agreement meant that the effect of not blinding and not randomly assorting did not significantly changed the error rate.

We analyzed the difference between random assortment and non-random assortment (for blind and not blind responses separately). In other words, we compared treatments 1 and 3 and compared treatments 2 and 4 to test the effects of not randomly assorting. We then analyzed the difference between blinding and no blinding (for random and not random responses separately). In other words, we compared treatments 1 and 2 and compared treatments 3 and 4 to test the effects of not blinding. We calculated significance using a two-proportion z-test, which tests the null hypothesis that two means are equal. Since we had four treatments for three question parts, we did 12 z-tests for each phase and coder. Significant alpha was 0.05. Becausewe did 12 tests per coder, the Bonferroni correction resulted in a significance threshold of 0.004167.

Bias

The methods that we used for this study on bias may have also introduced bias. For Phase I, only one coder completed the task. She coded a large number of responses for multiple question parts in order to minimize the effects of having only one coder. Additionally, although she randomly selected the order of treatments, since all responses of one treatment were coded together, the results may depict order effects and not necessarily treatment effects. We performed Phase II precisely for these reasons: Phase II used multiple coders and a split-plot design. Before

beginning Phase II, all coders knew of the results of Phase I (it was the results of Phase I that piqued interest in doing Phase II). Knowing the results may or may not impact our coding in Phase II, which will be discussed in the Discussion section.

Results

Phase I

During Phase 1, Coder X agreed with the computer codes significantly more often when responses were not blinded and when responses were in the order of the codes than when responses were blinded and randomly assorted ($p < 0.004167$; Figure 3). Of the 12 tests, two failed to show a significant difference: no effect of randomly assorting the responses for either blind ($z = 2.583$; $p = 0.009806$) or non-blind responses ($z = 2.658$; $p = 0.007861$) for the replication question part. The effect was significant for all transcription and translation tests.

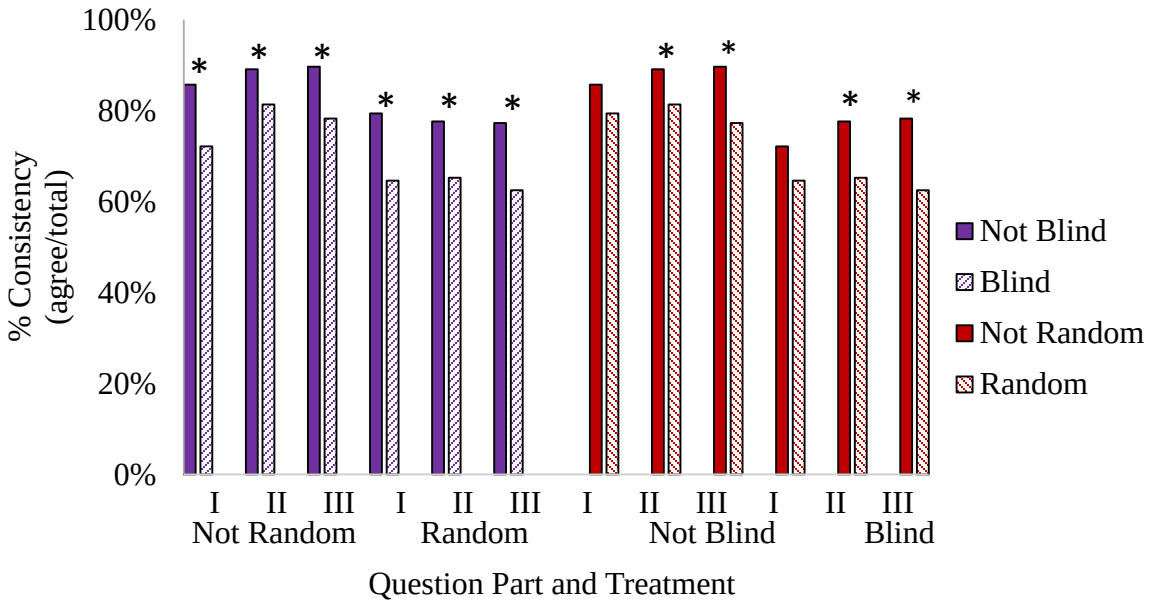

Figure 3. Phase I computer and human agreement for each question part (I = Replication, II = Transcription, and III = Translation) and treatment. * denotes statistical significance.

Phase II

Unlike Phase I, blinding the responses and randomly assorting them did not impact agreement for all three coders ($p > 0.004167$; Figures 4-6). Of the 36 tests, two exceptions occurred: Coder X agreed significantly more with the computer model when non-blinded responses for the replication part was randomly assorted rather than not-randomly assorted ($z = -4.321$; $p = 0.000016$), and Coder Y agreed significantly more with the computer model when randomly assorted responses for the translation part was not blinded rather than blinded ($z = 4.238$; $p = 0.000022$).

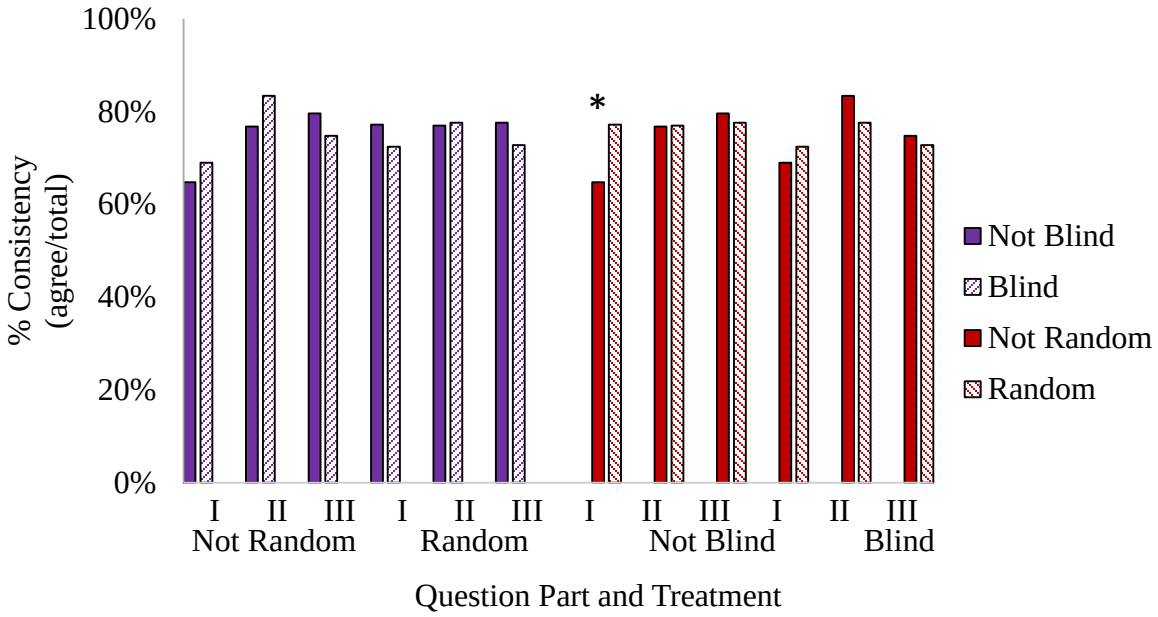

Figure 4. Phase II computer and human agreement for each question part (I = Replication, II = Transcription, and III = Translation) and treatment for Coder X

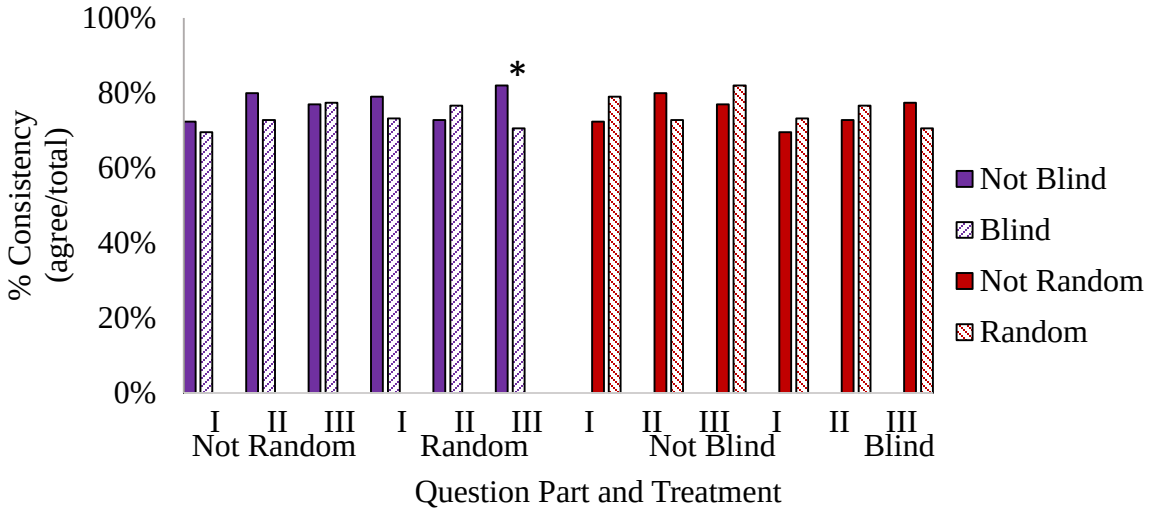

Figure 5. Phase II computer and human agreement for each question part (I = Replication, II = Transcription, and III = Translation) and treatment for Coder Y. * denotes statistical significance.

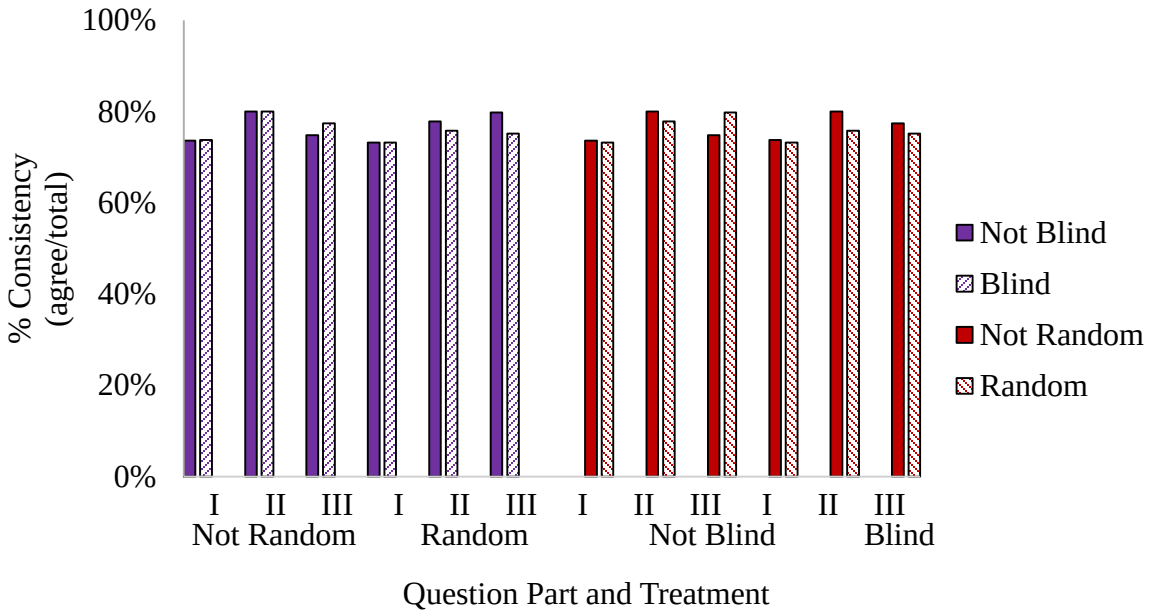

Figure 6. Phase II computer and human agreement for each question part (I = Replication, II = Transcription, and III = Translation) and treatment for Coder Z.

Discussion

In order to determine if we could improve computer models by verifying computer codes instead of coding blindly and randomly, we tested whether having the computer-predicted codes visible (i.e., no blinding) and having the responses in the order of the computer codes (i.e., no random assortment) created confirmation bias in human coding. When Coder X coded the 6,000 responses during Phase 1, she was confident that she would not be biased by having the computer codes visible. After obtaining evidence from Phase 1 that supported that she was biased in her coding, three coders, knowing the results of Phase 1, coded the same responses using the same treatments. This time, one of the coders was new to the project and Coders X and Y discovered that they needed to develop clearer definitions in the coding rubric in order to train her on the rubric. Additionally, this time we used a split-plot design (Figure 2). The results indicate that the three coders were not affected by having the codes visible and having responses in the order of the codes. Based on the methodological differences between Phases 1 and 2, we propose three possible reasons for these results: a) knowing about one's potential bias, b) having clearer coding definitions, and c) using a split-plot design may reduce confirmation bias.

It has been suggested that being aware of one's own potential bias can reduce bias (Bell and Mellor 2009; Rosenthal 1976). It is possible that its potential of reducing bias may depend on the type of data. While coding, some units may be more difficult to code than others. Coders may be less biased when coding takes little thinking compared to when coding takes more thought and time. In our study, during Phase 1, the coder believed that she would not be biased by having the codes visible, but results indicated that she was biased. Then during Phase 2, all three coders knew the results of Phase 1 and did not illustrate bias. By knowing about the potential bias, the coders may have been able to reduce their confirmation bias.

Coding rubrics are often tested using inter-coder reliability. Once inter-coder reliability is established, then the coding rubric is deemed reliable. However, the rubric may or may not be useful if the coders worked together on developing the rubric as inter-coder reliability really measures the coders' ability to code and does not test the rubric itself. There may be aspects of the codes that are understood by the developers but are not explicitly part of the coding rubric. These implicit aspects of the rubric may be forgotten or changed over time, and therefore, coding may become less consistent over time. In order to test a rubric, and not just the coders themselves, coders that are unfamiliar with the rubric but experts in the discipline should use it and inter-coder reliability should be established between coders unfamiliar and familiar with the rubric. In our study, once we invited a third coder on this project, we found that our rubric contained implicit aspects, and therefore, we had to define the codes in more detail. Similar to using different standards in clinical diagnostic tests may lead to biased results (Lijmer et al. 1999), having different versions of a rubric for the two phases may have impacted coding. It is possible that having a better defined rubric allows less room for questioning the codes and therefore, decreases confirmation bias.

A split-plot design allows mixing of treatments instead of coding all units of one treatment together. During coding, coders may experience fatigue or interpret a coding rubric differently over time. This error may be random, but if it does not occur equally across the treatments, then the error becomes systematic and will misrepresent the truth. Therefore, if treatments are not mixed at all, it is less reliable to state that any potential differences between treatments is due to the treatment and not fatigue or learning. Therefore, a split-plot design is a more reliable method. During Phase 1, all units of one treatment were coded together and then

during Phase 2, we used a split-plot design. In addition to reducing the effects of fatigue and learning, our results may support that a split-plot design may also reduce confirmation bias.

While Phase I illustrates the usefulness of blinding and random assortment, Phase II failed to show an effect. These results may indicate that blinding and random assortment are useful methods to use but may be unnecessary when implementing additional methods- such as being aware of one's own bias, testing the coding rubric with those that are unfamiliar with it, and using a split-plot design. These are encouraging results since during some research studies random assortment and blinding may not be possible. On the other hand, because every research project is so different and we do not think that a "one size fits all" will be possible, we encourage researchers to utilize as many methods as possible to reduce bias. Therefore, we still encourage the use of random assortment and blinding when possible.

As education researchers, we investigate an endless list of research questions to ultimately improve teaching and learning. Since educators may alter their practices based on research results, we need to make sure that these results are valid and reliable so that alterations lead to improvement. Therefore, we need to use as many methods as possible to reduce bias and increase reliability and validity of our research designs. By doing so, it will help us avoid misleading results and ultimately improve education.

Acknowledgements

This material is based upon work supported by the National Science Foundation (Grants 0736952, 0909999, 1022653, 1323162, and 1347740). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the supporting agencies.

References

Bell, I., and Mellor, D. 2009. Clinical Judgements: Research and Practice. *Australian Psychologist* 44(2): 112-121.

Bruine de Bruin, W. 2005. Save the Last Dance for Me: Unwanted Serial Position Effects in Jury Evaluations. *Acta Psychologica* 118(3): 245-260.

Butler, J.K., and Cantrell, R.S. 1986. Effects of Cue Order in a Decision-Modeling Instrument. *Psychological Reports* 58: 699-704.

Cabanac, G., and Reuss, T. 2013. Capitalizing on Order Effects in the Bids of Peer-Reviewed Conferences to Secure Reviews by Expert Referees. *Journal of the American Society for Information Science and Technology* 64(2): 405-415.

Chan, D.K.C., Ivarsson, A., Stenling, A., Yang, S.X., Chatzisarantis N.L.D., and Hagger, M.S. 2015. Response-Order Effects in Survey Methods: A Randomized Controlled Crossover Study in the Context of Sport Injury Prevention. *Journal of Sport & Exercise Psychology* 37: 666-673.

Guyatt, G., Furukawa, T. 2008. An Illustration of Bias in Random Error: Introduction. In: Guyatt, G., Rennie, D., Meade, M.O., Cook, D.J. (Eds.). *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice* (2nd ed.). Toronto: McGraw Hill: 109-112.

Hofer, S.I. 2015. Studying Gender Bias in Physics Grading: the Role of Teaching Experience and Country. *International Journal of Science Education* 37(17): 2879-2905.

Holman, L., Head, M.L., Lanfear, R., & Jennions, M.D. 2015. Evidence of Experimental Bias in the Life Sciences: Why We Need Blind Data Recording. *PLOS Biology* 13(7): e1002190.

Jones, B., and Nachtsheim, C.J. 2009. Split-Plot Designs: What, Why and How. *Journal of Quality Technology* 41(4): 340-361.

Landis, J.R., and Kock, G.G. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33: 159-174.

Lijmer, J.G., Mol, B.W., Heisterkamp, S., Bonsel, G.J., Prins, M.H., van der Meulen, J.H.P., and Bossuyt, P.M.M. Empirical Evidence of Design-Related Bias in Studies of Diagnostic Tests. *The Journal of the American Medical Association* 282(11): 1061-1063.

MacCoun, R.J. 1998. Biases in the Interpretation and Use of Research Results. *Annual Review of Psychology* 49: 259-287.

Malouff, J.M., Emmerton, A.J., and Schutte, N.S. 2013. The Risk of a Halo Bias as a Reason to Keep Students Anonymous During Grading. *Teaching of Psychology* 40(3): 233-237.

Mantonakis, A., Rodero, P., Lesschaeve, I., and Hastie, R. 2009. Order in Choice: Effects of Serial Position on Preferences. *Psychological Science* 20(11): 1309-1312.

McClendon, M.J. 1986. Response-Order Effects for Dichotomous Questions. *Social Science Quarterly* 67: 205-211.

Newman, J.L., and Fuqua, D.R. 1992. Effects of Order of Presentation on Perceptions of the Counselor. *Journal of Counseling Psychology* 39(4): 550-554.

Rosenthal, R. 1976. *Experimenter Effects in Behavioral Research*. New York: Irvington Publishers, Inc.

Rosenthal, R., and Jacobson, L. 1968. Pygmalion in the Classroom. *The Urban Review* 3(1): 16-20.

Smith, M.K., Wood, W.B., and Knight, J. 2008. The Genetics Concept Assessment: A New Concept Inventory for Gauging Student Understanding of Genetics. *CBE: Life Sciences Education* 7: 422-430.

Urban-Lurain, M., Cooper, M.M., Haudek, K.C., Kaplan, J.J., Knight, J.K., Lemons, P.P., Lira, C.T. et al. 2015. Expanding a National Network for Automated Analysis of Constructed Response Assessments to Reveal Student Thinking in STEM. *Computers in Education Journal* 6: 65-81.

Willits, F.K., and Saltiel, J. 1995. Question Order Effects on Subjective Measures of Quality of Life. *Rural Sociology* 60(4): 654-665.

Yates, F. 1935. Complex Experiments. *Supplement to the Journal of the Royal Statistical Society* 2(2): 181-247.