Royal, Kenneth D. 2016. "The Impact of Item Sequence Order on Local Item Dependence: An Item Response Theory Perspective." *Survey Practice* 9 (5). https://doi.org/10.29115/SP-2016-0027.

ARTICLES

The Impact of Item Sequence Order on Local Item Dependence: An Item Response Theory Perspective

Kenneth D. Royal[1]

[1] North Carolina State University

Keywords: bias, item order, best practice, psychometrics, statistical dependency, local item dependence, rasch measurement, survey research

https://doi.org/10.29115/SP-2016-0027

Survey Practice

Vol. 9, Issue 5, 2016

Local item dependence (LID) suggests a response to one item was directly influenced by a response to another item. Items exhibiting LID typically cause survey participants to provide biased/inaccurate responses, which ultimately poses a threat to score validity. Further, conflicting 'best practice' guidelines from survey research experts on how to construct surveys, particularly whether related items should be presented in a random or consistent order, may contribute to LID bias. The purpose of this paper is to bring attention to the issue of LID via a case example/experiment and illustrate how following best practice guidelines for survey construction may actually increase bias/error in some instances. Implications and recommendations are discussed.

Introduction

Survey researchers have used item response theory (IRT) models for decades to evaluate the psychometric properties of survey instruments and analyze subsequent data. The rationale for using these models typically involves the researcher's ability to overcome well-documented problems and limitations pertaining to the use of raw scores. Royal (2010) notes six specific weaknesses and limitations of traditional statistical analyses: (1) most survey scales (e.g. Likert-type scales, visual analog scales, semantic differential scales, etc.) are ordinal in nature but are erroneously treated as interval level; (2) items are assumed to be of equal difficulty and merit; (3) error estimates are assumed to be the same for all participants; (4) results are sample-dependent and inherently linked specifically to the participants' that completed the survey; (5) parametric statistical approaches require normally distributed data; and (6) missing data are often a problem for statistical analyses that do not use some variation of a maximum likelihood estimation. Over the last decade, in particular, advances in statistical and measurement software programs have made it more feasible for nonspecialists to use IRT models; thus, the use of IRT for survey validation, quality control, and data analysis has become increasingly popular and quite common in modern research.

Although the family of IRT models is broad, one of the most commonly used models for survey research is the Rasch Rating Scale Model (Andrich 1978). Many scholars consider this model (and Rasch models in general) to be among the gold standard approaches for data analysis (Bond and Fox 2015) because Rasch models overcome the previously noted limitations and weakness of traditional statistical approaches and are invariant when data sufficiently

fit the model. In short, Rasch models are probabilistic models that converts ordinal data to linear measures using the logarithmic scale. Two parameters are then modeled against one another, a latent trait for the participants, and an estimate of difficulty for each of the items. Fit statistics evaluate response patterns and identify persons and items that misfit the model's expectations and introduce 'noise' into the measurement system. Because misfitting data distorts objective measurement, data that misfit given context-specific ranges (e.g. test, surveys, performance assessments, etc.) (Wright and Linacre 1994) are often discarded. In the case of survey research, latent traits modeled typically include factors such as the strengths of one's attitude or preference, the tendency for a participant to agree (or disagree) with a given item, and so on. When modeled, the likelihood that a given person would endorse an item is a logistic function of the distance between the participant and the item on the continuum. Although a thorough discussion of Rasch models is beyond the scope of this paper, readers are encouraged to see Engelhard (2013) and Bond and Fox (2015) for a thorough overview.

A requirement for objective measurement is local independence of items. That is, when items demonstrate local item dependence (LID), also referred to as statistical dependency, it suggests a response to one item was directly influenced by a response to another item (Marais and Andrich 2008). Items that are locally dependent typically cause participants to provide bias/inaccurate responses, so the implications regarding validity evidence is quite significant. Survey researchers using IRT models routinely investigate LID when evaluating the psychometric properties of an instrument. Typically, when statistically dependent items are discovered, they are reviewed for content, and a decision is made to either retain, revise, or discard one or more of the potentially dependent items.

At present, there is a significant lack of survey research literature that addresses the problem of LID, particularly as it pertains to identifying LID with item response theory models and treating locally dependent items. Further, conflicting guidelines from survey research experts on how to construct surveys, particularly whether items should be presented in a random or consistent order, may contribute to LID bias that ultimately affects both the accuracy of the results, and the appropriateness of the inferences that are made about the results. Thus, the purpose of this paper is to bring attention to the issue of LID via a case example/experiment and illustrate how following best practice guidelines for survey construction may actually increase bias/error in some instances.

Background and Context

In public opinion studies, surveys are often very long, so researchers routinely randomize the order in which items are presented to respondents as a way to ensure response coverage (e.g. some responses are collected for all items, as opposed to many responses for only those items appearing at the beginning

of an instrument) should survey fatigue and/or attrition be a concern. Randomization also works well when many constructs and/or topics are presented and subsets of items can be randomized to ensure response coverage. However, in other areas of research and practice, surveys are often much shorter and many, if not most, of the items pertain to a single subject matter. For example, healthcare professionals might administer depression inventories, functional mobility scales, pain scales, and so on. Market researchers might administer surveys measuring consumer satisfaction, television viewing habits, product preferences, and so on. Thus, in these contexts, the need for randomization might be less than that of larger surveys.

As noted previously, there appears to be some confusion in the survey research literature as published guidelines regarding best practices for survey construction often vary. For example, many survey researchers consider it a best practice to group related items together as it makes it easier for participants to complete the survey, gives the appearance of greater cohesiveness, and requires a lesser cognitive load from participants (Bradburn, Sudman, and Wansink 2004; Dillman 2000). Yet, a considerable body of extant research has suggested the exact opposite should occur. More specifically, item order should be randomized when possible because item order effects, often called assimilation effects or carry-over effects, can result in biased participant responses (Heiman 2002). It is this researcher's opinion that these conflicting guidelines are due to contextual differences and assumptions made about the types of surveys administered. Nonetheless, the problem remains quite severe, as many survey researchers acquire best practices from a variety of sources, many of which are in discipline-specific areas, and the consequences that may result from erroneous interpretations of context-free information could pose a significant validity threat for many studies.

Case Example

An academic misconduct survey was administered to doctor of veterinary medicine program students at a large college of veterinary medicine in the United States (Royal and Flammer 2015; Royal, Schoenfeld-Tacher, and Flammer 2016). The survey contained 23 items measuring the extent to which various actions and behaviors constitute academic misconduct. A 7-point semantical differential scale (1=Not Misconduct to 7=Severe Misconduct) was used to capture participants' perspectives. A total of 137 students completed the survey.

As part of the routine psychometric analysis, statistical dependence was investigated by conducting a Rasch-based principle components analysis (PCA) of standardized residuals (Linacre 2016a) and reviewing the residual item correlation matrix. Items with residual correlations greater than 0.3 were considered to exhibit LID (Smith 2000). Results of the psychometric investigation indicated four pairs of items exhibited LID (see Table 1), and interestingly, each pair of items were presented next to one another on the

Table 1 Statistically dependent item pairs based on veterinary students' responses.

Item	Correlation
#1: Copying from another student during a quiz or exam	0.69
#2: Using unauthorized cheat sheets or other materials during a quiz or exam	
#12: Missing class or lab due to a false excuse	0.37
#13: Claiming to have attended class when you actually did not	
#19: Failing to prepare adequately for a group assignment or laboratory	0.67
#20: Doing less than your fair share in a group project or a laboratory	
#22: Presenting your clinical skills book for signing without actually completing the skill	0.60
#23: Listing false completions on your online clinical skills completion summary	

instrument. A simple experiment was conducted to determine if the items exhibiting LID were truly dependent and likely to lead to biased responses, or if item ordering effects caused these items to exhibit LID.

Experiment

Two questionnaires were created with one version containing the items in the same order as originally presented (control) and the other version containing items presented in random order (experiment). The survey was presented to workers on Amazon Mechanical Turk, a crowdsourcing Internet marketplace available through Amazon.com. Qualtrics survey software was used to administer the survey. The first 100 participants completing each survey was compensated with a small stipend for their time and effort. Individuals who participated in one survey were ineligible for participation in the other, ensuring 200 distinct individuals from the same population group completed the surveys. Permission to conduct this study was granted by the North Carolina State University Institutional Review Board that declared the study exempt.

Quality Control and Data Analysis

Upon data collection, a series of routine quality control checks were performed as part of the initial Rasch measurement analysis. Data from both sets were evidenced to be mostly unidimensional, highly reproducible (Cronbach's >0.90), and fit the Rasch Rating Scale Model quite well. In order to obtain excellent data to model fit, misfitting data with Infit or Outfit mean square values greater than 2.0 (Wright and Linacre 1994) were removed as these data contributed noise (error) that distorted the measurement system. In total, 10 misfitting persons were removed from the control group, and two misfitting persons were removed from the experimental group. Winsteps (Linacre 2016b) measurement software was used to perform the IRT analysis.

Results

Item pairs that were previously flagged as potentially statistically dependent based on their residual correlations were investigated in both the control and experimental data sets (see Tables 2 and 3).

Table 2 Residual correlations based on control group responses.

Item	Correlation
#1: Copying from another student during a quiz or exam	0.40
#2: Using unauthorized cheat sheets or other materials during a quiz or exam	
#12: Missing class or lab due to a false excuse	0.32
#13: Claiming to have attended class when you actually did not	
#19: Failing to prepare adequately for a group assignment or laboratory	0.68
#20: Doing less than your fair share in a group project or a laboratory	
#22: Presenting your clinical skills book for signing without actually completing the skill	0.56
#23: Listing false completions on your online clinical skills completion summary	

Table 3 Residual correlations based on experimental group responses.

Item	Correlation
#1: Copying from another student during a quiz or exam	0.32
#2: Using unauthorized cheat sheets or other materials during a quiz or exam	
#12: Missing class or lab due to a false excuse	0.08
#13: Claiming to have attended class when you actually did not	
#19: Failing to prepare adequately for a group assignment or laboratory	0.32
#20: Doing less than your fair share in a group project or a laboratory	
#22: Presenting your clinical skills book for signing without actually completing the skill	0.28
#23: Listing false completions on your online clinical skills completion summary	

Results from the control study indicated that when the items were presented in the same order as the items originally presented on the veterinary student survey, each pair of items was once again flagged as exhibiting LID. However, when the items were randomly presented to participants in the experimental group, evidence of LID was greatly reduced. In fact, two pairs of items fell below the suggested threshold of 0.30, and the remaining two pairs fell to 0.32 (just slightly above the suggested threshold). Based on this evidence, it was clear that item ordering impacted participants' responses on this survey, and these items exhibited LID largely because the items were presented in close proximity.

Discussion and Recommendations

Many survey researchers and practitioners routinely revise or discard one or more survey items that are flagged as statistically dependent. The results of this experiment suggest one should use additional caution before altering or removing item. More specifically, evidence of LID may not necessarily be attributed to substantive items that are similar/related, but instead attributed to the order and proximity in which they were presented to participants. If a survey researcher rushes to judgment, perfectly good items may be subject to revision or discarded completely.

It is clear from this experiment that randomizing items may reduce biased responses resulting from LID contamination. Thus, survey researchers perhaps should employ a randomized item order when possible, provided there is no reason (philosophical or otherwise) not to do so. As noted previously, however, the context in which surveys are administered matters a great deal and likely is the reason why conflicting best practice guidelines sometimes exist. To that end, it is important to note that randomization is often not possible for many surveys, particularly those administered via paper-and-pencil which are incredibly common across the spectrum of healthcare fields. So, what can survey researchers do when item order cannot be randomized or when there is some compelling reason not to randomize? Survey researchers have several options, but perhaps the two simplest solutions include: (1) creating multiple forms of a survey in which the item order is different for each and (2) analyze pilot test data and investigate the presence of LID. If items exhibit LID during pilot testing, then this would be an appropriate time to either re-order items and administer a second pilot study or possibly revise/discard items when a qualitative review indicates just cause.

In closing, it cannot be emphasized enough that there are thousands of survey researchers throughout the world, and not all of them read the same literature. For example, researchers will often seek literature in their own familiar field, as opposed to searching for potentially more authoritative sources in other arenas (Royal and Rinaldo 2016). In fact, many researchers may be completely unaware that there is even a field of survey research as a formal field of inquiry. In other instances, survey researchers may only seek out best practices for surveys relating to a specific context (e.g. online surveys, surveys administered to college students, and so on). Regardless, it important to note that sometimes best practices translate smoothly across disciplines and contexts, but sometimes they do not. In any instance, there is no substitute for careful planning, pilot testing, quality assurance, thorough analyses, and critical interpretation of results.

REFERENCES

Andrich, D. 1978. "A Rating Formulation for Ordered Response Categories." *Psychometrika* 43 (4): 561–73.

Bond, T.G., and C.M. Fox. 2015. *Applying the Rasch Model. Fundamental Measurement in the Human Sciences*. 3rd ed. New York: Routledge.

Bradburn, N., S. Sudman, and B. Wansink. 2004. *Asking Questions: The Definitive Guide to Questionnaire Design*. San Francisco: Jossey-Bass.

Dillman, D. 2000. *Mail and Internet Surveys*. New York: John Wiley & Sons.

Engelhard, G. 2013. *Invariant Measurement: Rasch Measurement in the Social, Behavioral and Health Sciences*. New York: Routledge.

Heiman, G.W. 2002. *Research Methods in Psychology*. 3rd ed. Boston and New York: Houghton Mifflin Company.

Linacre, J.M. 2016a. "Dimensionality: Contrasts & Variances." 2016. http://www.winsteps.com/winman/principalcomponents.htm.

———. 2016b. "WINSTEPS® (Version 3.92.0)." Computer Software. 2016.

Marais, I., and D. Andrich. 2008. "Effects of Varying Magnitude and Patterns of Response Dependence in the Unidimensional Rasch Model." *Journal of Applied Measurement* 9 (2): 105–24.

Royal, K.D. 2010. "Making Meaningful Measurement in Survey Research: A Demonstration of the Utility of the Rasch Model." *IR Applications* 28: 1–16.

Royal, K.D., and K. Flammer. 2015. "Measuring Academic Misconduct: Evaluating the Construct Validity of the Exams and Assignments Scale." *American Journal of Applied Psychology* 4 (3–1): 58–64.

Royal, K.D., and J.C.B. Rinaldo. 2016. "There's Education, and Then There's Education in Medicine." *Journal of Advances in Medical Education and Professionalism* 4 (3): 150–54.

Royal, K.D., R. Schoenfeld-Tacher, and K. Flammer. 2016. "Comparing Veterinary Student and Faculty Perceptions of Academic Misconduct." *International Research in Higher Education* 1 (1): 81–90.

Smith, R.M. 2000. "Fit Analysis in Latent Trait Measurement Models." *Journal of Applied Measurement* 1 (2): 199–218.

Wright, B.D., and J.M. Linacre. 1994. "Reasonable Mean-Square Fit Values." *Rasch Measurement Transactions* 8 (3): 370.