Human Resource Management Review xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Human Resource Management Review

journal homepage: www.elsevier.com/locate/humres

Exploratory data analysis as a foundation of inductive research

Andrew T. Jebb *, Scott Parrigon, Sang Eun Woo

Department of Psychological Sciences, Purdue University, 703 Third Street, West Lafayette, IN 47907, United States

ARTICLE INFO

Available online xxxx

Keywords:
Exploratory data analysis
Inductive
Phenomenon detection
Replication
Visualization

ABSTRACT

Across academic disciplines, scientific progress is maximized when there is a balance between deductive and inductive approaches. To promote this balance in organizational science, rigorous inductive research aimed at phenomenon detection must be further encouraged. To this end, the present article discusses the logic and methods of *exploratory data analysis* (EDA), the mode of analysis concerned with discovery, exploration, and empirically detecting phenomena in data. We begin by first describing the historical and conceptual background of EDA. We then discuss two issues related to EDA and its relationship to scientific credibility. First, we argue that EDA fosters a replication-based science by requiring cross-validation and by emphasizing the natural uncertainty of data patterns. Second, we clarify that EDA is distinguishable from other exploratory practices that are considered scientifically questionable (e.g., "*p*-hacking", "data fishing" and "data-dredging"). In the following section of the paper, we present a final argument for EDA: that it helps maximize the value of data. To illustrate this point, we present several graphical methods for detecting data patterns and provide references to further techniques for the interested reader.

In organizational science, a great deal of emphasis is placed on the importance of causal explanations, or *theories* (Hambrick, 2007; Sutton & Staw, 1995). Out of these explanations, we gain an understanding of events in the world, an understanding which allows for their prediction and control. Thus, good theories are vital, and they derive their value from the *phenomena*, or "robust empirical regularities" they describe (Haig, 2005, p. 372). The starting point of theory can, therefore, be said to lie in first identifying phenomena to explain. This scientific activity, *phenomenon detection*, is fundamental to scientific progress. Logically, it precedes theory development, refinement, and testing, and its importance is reflected in the fact that "more Nobel prizes are awarded for the discovery of phenomena than for the construction of explanatory theories" (Haig, 2005, p. 384). Thus, theory may be the overarching "story," but phenomena are the ever-important words with which it is written.

The objective of the current article is to introduce *exploratory data analysis* for more effective phenomenon detection within organizational science. We believe this is sorely needed. Although inductive approaches have become more appreciated in the field (Spector, Rogelberg, Ryan, Schmitt, & Zedeck, 2014), the bulk of organizational research remains strongly oriented toward hypothetico-deductive confirmatory studies. In these settings, the entire domain of theory is specified a priori, and there is little room for discovery. It is entirely possible that the enterprise of phenomenon detection has suffered as a result (Hambrick, 2007).

There is also a lack of published material on exploratory data analysis (EDA) within the methodological literature. In our review of the organizational sciences literature, we found no published papers on EDA, other than a short book review of Tukey's (1977), *Exploratory Data Analysis*, published in *Administrative Science Quarterly* (Mueller, 1980). This lack of formal resources to

* Corresponding author.
E-mail addresses: ajebb@purdue.edu (A.T. Jebb), sparrigo@purdue.edu (S. Parrigon), sewoo@purdue.edu (S. Eun Woo).

http://dx.doi.org/10.1016/j.hrmr.2016.08.003
1053-4822/

Please cite this article as: Jebb, A.T., et al., Exploratory data analysis as a foundation of inductive research, *Human Resource Management Review* (2016), http://dx.doi.org/10.1016/j.hrmr.2016.08.003

guide EDA may inhibit proper channels for data exploration in discovering important workplace phenomena. It may also contribute to issues of scientific (un)reliability. As noted by many scholars (e.g., Kepes & McDaniel, 2013), researchers may conduct analyses that contain exploratory elements but then package them within a final confirmatory product. This mixing of exploratory behaviors within confirmatory settings allows the data to simultaneously generate and test the analytic plan, leading to hypothesizing after the results are known (HARKing; Kerr, 1998) and immunizing scientific hypotheses from falsification. We believe that providing a formal discussion on what EDA *is*, what it is not, and its relationship to confirmatory research, can help alleviate certain issues related to research integrity. Our paper takes an initial step toward this purpose.

We begin our paper with a formal introduction to EDA and a description of its history and underlying philosophy. The remainder of the paper focuses on theoretical issues related to the role of EDA within organizational science and is divided into two major sections. The first addresses the relationship between EDA and research credibility. We discuss how EDA, when properly understood and practiced, *encourages* scientific replication, and how it is entirely removed from other forms of exploratory statistics that are scientifically questionable (e.g., "fishing," "data snooping," and "*p*-hacking"). While concerns regarding research integrity should be properly raised and addressed, EDA might be mistakenly tied into these practices, causing researchers to feel hesitant when conducting exploratory analyses or simply to avoid them altogether. Therefore, we set our first task as defining the hard, *uncompromising* line between confirmatory data analysis (CDA) and EDA, whose "fuzziness" has, for too long, allowed the exploratory to slip into the confirmatory.

The second major section highlights one of the chief benefits of EDA: that it *maximizes the value of the data*. The scientific goals of CDA constrain it to make minimal use of the data. Confirmatory research is desirable for testing and/or validating specific effects that are theoretically expected, but without exploratory work, a great deal of data goes to waste. Thus, EDA is a necessary adjunct for organizational scholars to get the most from their collected data. After discussing these points conceptually, we then present several graphical techniques for phenomenon detection and provide further references for the interested reader.

In discussion of these issues, we hope to invite a certain degree of change within organizational research practices. There is much to gain from regular and explicit exploratory work: a more reliable literature, valuable scientific insights, better designed confirmatory studies, and a fuller understanding of organizational phenomena. Put more accurately, there is much to gain from realizing the degree to which exploratory research is the natural *complement* to confirmatory analysis.

1. A brief introduction to EDA

Before providing an outright definition of EDA, it is helpful to first describe its historical and ideological roots. EDA largely developed out of the pioneering work of renowned statistician John Tukey and colleagues, presented in his classic work, *Exploratory Data Analysis* (Tukey, 1977). Though certain EDA techniques existed prior to Tukey (Leinhardt & Leinhardt, 1980), this landmark volume consolidated its underlying ideals and techniques within a single structured framework. Within its pages, Tukey argued that far too much emphasis had been placed on confirmatory research and far too little emphasis on using data to *develop* theory, suggest research designs, and ensure proper confirmatory analyses (Tukey, 1980). Throughout his career, Tukey repeatedly stressed the *complementary* relationship between EDA and CDA and thus between inductive and deductive approaches (Tukey, 1980). Replacing "either by the other is madness," he wrote (1980, p. 23). On the statistical side, EDA ensures that confirmatory models are accurately specified and that their assumptions are actually borne out in the data. On the theoretical level, EDA leads to better articulated research questions, stronger research hypotheses, and finding unexpected phenomena that could not have been deduced a priori. This inductive-deductive synergy, a hallmark of strong scientific work, is represented statistically within the joint use of EDA and CDA.

Due to the importance of the exploratory mode of analysis, other expositions soon followed Tukey (1977), including Mosteller and Tukey (1977), McNeil (1977), and Erickson and Nosanchuk (1977; see Behrens, 1997, p. 132 for others). This methodology has also seen many recent refinements, especially in light of advances in computing power and data visualization. Several of the most prominent in this class are Cleveland's (1994) *Visualizing Data*, Theus and Urbanek's (2008) *Interactive Graphics for Data Analysis: Principles and Examples*, and Edward Tufte's (1983) *The Visual Display of Quantitative Information*.

Being multifaceted, a single definition of EDA is elusive. EDA is not a type of model or even a defined set of procedures. EDA in action is also highly context-specific; depending on the details of the analysis, EDA in one application may look entirely different from another. This is because unlike CDA, EDA is characterized by an extreme flexibility that is necessary for identifying and investigating the range of statistical and substantive phenomena that emerge during empirical research. EDA has been characterized in different ways which underscore its many aspects. In a series of lectures, Tukey (1993) noted many of the diverse principles that underlie EDA, such as a willingness to find both unexpected and regular phenomena, a flexible mental attitude, smoothing techniques, simple mathematical calculations, *seeing* the results (i.e., visualization), robust statistics, a view of data as *fit* + *residual*, model *building*, the incompleteness of all models, and that patterns need not be "large" to be noteworthy.

Another viewpoint is given by Hartwig and Dearing (1979), who describe the "exploratory perspective" as comprising two parts: *openness* and *skepticism* (p. 9). Openness is a readiness to find patterns that are different than expected. Such patterns may signal either (a) novel theoretical insights or (b) information to guide further analyses (e.g., the need for specific covariates or a multilevel model). Openness also relates to variable *re-expression*. The original scales of variables are often arbitrary or a matter of convention and convenience rather than having substantive meaning (Behrens, 1997; Hartwig & Dearing, 1979). Exploring scale transformations through EDA can facilitate interpretation, standardize dispersion across groups, and ensure statistical assumptions are met (e.g., linearity; Emerson & Stoto, 1983).

Please cite this article as: Jebb, A.T., et al., Exploratory data analysis as a foundation of inductive research, *Human Resource Management Review* (2016), http://dx.doi.org/10.1016/j.hrmr.2016.08.003