Journal of Management
Vol. 39 No. 7, November 2013 1799-1824
DOI: 10.1177/0149206312466147

Using Market Basket Analysis in Management Research

Herman Aguinis
Lura E. Forcum
Harry Joo
Indiana University

Market basket analysis (MBA), also known as association rule mining or affinity analysis, is a data-mining technique that originated in the field of marketing and more recently has been used effectively in other fields, such as bioinformatics, nuclear science, pharmacoepidemiology, immunology, and geophysics. The goal of MBA is to identify relationships (i.e., association rules) between groups of products, items, or categories. We describe MBA and explain that it allows for inductive theorizing; can address contingency (i.e., moderated) relationships; does not rely on assumptions such as linearity, normality, and residual equal variance, which are often violated when using general linear model–based techniques; allows for the use of data often considered "unusable" and "messy" in management research (e.g., data not collected specifically for research purposes); can help build dynamic theories (i.e., theories that consider the role of time explicitly); is suited to examine relationships across levels of analysis; and is practitioner friendly. We explain how the adoption of MBA is likely to help bridge the much-lamented micro–macro and science–practice divides. We also illustrate that use of MBA can lead to insights in substantive management domains, such as human resource management (e.g., employee benefits), organizational behavior (e.g., dysfunctional employee behavior), entrepreneurship (e.g., entrepreneurs' identities), and strategic management (e.g., corporate social responsibility). We hope our article will serve as a catalyst for the adoption of MBA as a novel methodological approach in management research.

Keywords: *research design; methodology; data analysis*

Acknowledgments: We thank J. Craig Wallace, two Journal of Management *anonymous reviewers, Jose M. Cortina, Brian K. Boyd, Steven E. Markham, and Sofia J. Vaschetto for detailed and highly constructive feedback on previous drafts. A previous version of this manuscript was presented at the annual meetings of the Academy of Management, Boston, Massachusetts, August 2012.*

Corresponding author: Herman Aguinis, Department of Management and Entrepreneurship, Kelley School of Business, Indiana University, 1309 E. 10th Street, Bloomington, IN 47405-1701, USA

E-mail: haguinis@indiana.edu

The goal of our article is to describe market basket analysis (MBA), which is a methodological approach new to the field of management, and to explain and illustrate that its adoption is likely to lead to substantive theoretical advancements as well as meaningful practical applications. Our goal addresses recent calls for expanded and improved methodological tools (Aguinis, Dalton, Bosco, Pierce, & Dalton, 2011; Bowen, 2012; Boyd, Haynes, Hitt, Bergh, & Ketchen, 2012; Ployhart & Vandenberg, 2010).

MBA, also known as association rule mining or affinity analysis, is a data-mining technique that originated in the field of marketing to identify relationships between groups of products, items, or categories. Besides its specific advantages for management research, which we describe in a later section of our article, MBA has the potential to play an important role in bridging two much-lamented divides. First, MBA can help bridge the micro (e.g., organizational behavior, human resource management, industrial and organizational psychology) versus macro (e.g., entrepreneurship, strategy, organizational theory) divide (Aguinis, Boyd, Pierce, & Short, 2011). Micro and macro researchers tend to have different areas of interest, write different types of articles, publish in different journals, and even use different methods in their research (Aguinis, Boyd, et al., 2011). MBA can be used to address research questions relevant to both micro and macro research domains and understand relationships between microlevel variables (e.g., individual knowledge and skills) and macrolevel variables (e.g., organizational policies). Moreover, MBA can be used inductively to generate hypotheses and build theories that integrate and synthesize micro- and macrolevel variables. Second, MBA can help bridge the science–practice divide. As noted by Cascio and Aguinis (2008: 1062), "there is a serious disconnect between the knowledge that academics are producing and the knowledge that practitioners are consuming." As we describe in the following, MBA can be used to address not only questions that are highly relevant for theory development but also questions of importance for practitioners—as has been done in the sister field of marketing. Stated differently, the use of MBA can produce knowledge that is relevant and actionable, and it therefore can play an important role in helping bridge the science–practice divide.

The remainder of our article is organized as follows. First, we describe MBA and provide examples of its use in fields outside of management. Second, we explain the advantages of using MBA in management research. Third, we provide a step-by-step explanation of how to use MBA. This description is sufficiently detailed such that researchers completely unfamiliar with MBA will be able to learn its fundamental features. Fourth, we illustrate how MBA can be used to investigate substantive questions in various management domains. Finally, we conclude with a description of potential pitfalls to avoid when using MBA.

MBA

MBA allows researchers to uncover nonobvious and usually hidden and counterintuitive associations between products, items, or categories. This methodological approach allows researchers to identify those items that co-occur (i.e., appear together) on a frequent basis and assess the extent to which they co-occur. MBA has been used to understand consumer behavior regarding types of books that are purchased together (as purchased on Amazon

.com) as well as different types of wines that the same individual is likely to purchase (as purchased on VirginWines.com) (Berry & Linoff, 2004). Because MBA originated in the field of marketing and was initially used to understand which supermarket items are purchased together (i.e., placed together in the same "basket"), the technique adopted the name *market basket analysis*.

Agrawal, Imieliński, and Swami (1993) seem to have used MBA first (Nisbet, Elder, & Miner, 2009). Agrawal et al. are computer scientists who had access to a large repository of previously collected customer transaction data and were able to discover association rules between items purchased. The method was quickly adopted in the field of marketing as a popular tool for a variety of practical applications. For example, suppose that a marketing researcher uses MBA to gather empirical support that cake mix and icing tend to be purchased together. The researcher may then determine that cake mix and icing are complementary items (i.e., categories) such that lowering the price of only one of the goods is associated with an increase in demand for both goods (Y.-L. Chen, Tang, Shen, & Hu, 2005). From a theoretical standpoint, MBA can be used to develop hypotheses and theories inductively. For example, these results suggest the hypothesis that consumers may have mental models that include associations among many other sets of items that are complementary in terms of their activities and interests (e.g., running shoes and water bottles). From a practical standpoint, this study's results can be used to make decisions such as stocking the two items near each other, thereby increasing the likelihood that customers will easily find and purchase both products instead of just one of them (Russell & Petersen, 2000).

The applicability of MBA is not limited to analyzing archival data. The method can also be used with primary data. For example, Kanagawa, Matsumoto, Koike, and Imamura (2009) mailed out surveys in which respondents completed open-choice checklists regarding common food allergens. As a result, Kanagawa et al. found that certain food allergens tend to occur together in the same person. As a second illustration of the use of MBA with primary data, Goh and Ang (2007) collected responses from secondary school students in Singapore about their preferences regarding the use of school-based counseling services. On the basis of their responses to the administered questionnaires, students were separated into two groups—those willing to see a counselor and those unwilling—and Goh and Ang identified the main traits and concerns of each group. In short, although MBA has traditionally been used with archival data, it can also be used with primary data.

Categorical association rules are based on binary data, and this is the most common type of association rule because MBA was originally developed to analyze shopping cart data in binary form: Either a customer bought or did not buy a particular product. However, it is possible to derive association rules involving continuous variables (e.g., Aumann & Lindell, 2003; Bennedich, 2009). Association rules involving a continuous variable are called quantitative association rules. Categorical association rules relate a value of a categorical variable with a value of another categorical variable, but a quantitative association rule relates a value of a categorical variable with a summary statistic of a continuous variable (e.g., mean, median). As an example, Bennedich (2009) used MBA to derive quantitative association rules based on hotel customers. The data set included responses to Likert-type scales assessing behavioral and attitudinal constructs provided by customers of a large hotel chain. Bennedich derived several quantitative association rules, such as those relating certain floor

numbers with mean room satisfaction scores. In short, MBA allows for the derivation of association rules including not only categorical but also continuous variables, such as those measured using multiple-item scales.

Understanding the Presence and Strength of Association Rules

Three indexes are commonly used to understand the presence, nature, and strength of an association rule: lift, support, and confidence (Berry & Linoff, 2004; Larose, 2005; Zhang & Zhang, 2002). All three indexes are used because they provide complementary and non-redundant information. Lift is obtained first because it provides information on whether an association actually exists as well as whether the association is positive or negative. If the value for lift suggests that an association rule exists, the next step is to obtain the value for support, which is the actual probability that a set of items co-occurs with another set of items in a data set. Then, confidence is computed, which is the probability that a set of items occurs given that another set of items has already occurred.

Lift is defined as $\frac{P(A \cap B)}{P(A) * P(B)}$. The denominator assumes that Events A and B occur independently of each other (i.e., have no association), expressed by the multiplication of the probability of A with that of B. On the other hand, the numerator assumes that A and B co-occur (i.e., are associated), an assumption that is reflected in the probability of the union of the two events. Thus, if the numerator is similar to the denominator, lift will be close to 1.0, indicating that the relationship between A and B can be explained by chance. For this reason, lift is used as a means of screening out association rules whose lift values are equal or close to 1.0 (Baralis, Cagliero, Cerquitelli, Garza, & Marchetti, 2011). Note that lift will be greater than 1.0 when the presence of A is associated with the presence of B—a relationship that is positive in nature. On the other hand, lift is lower than 1.0 when the numerator (assuming a positive relationship) is smaller than the denominator (assuming no relationship). In such cases, lift indicates a relationship that is negative in nature: The presence of A is associated with the absence of B.

Lift is obtained first because it serves a function similar to statistical significance testing in more traditional analyses. That is, lift provides a decision-making tool to conclude whether the presence of an association cannot be explained by chance alone. Although there are ways to assess the statistical significance of association rules (e.g., Alvarez, 2003; Webb, 1999), they are not very informative and, instead, lift is used widely (Berry & Linoff, 2004). This choice is understandable because, given the large sample sizes used in MBA studies, there would be few associations found not to be statistically significant. However, a caveat regarding the use of lift is that it requires a judgment call on the part of the user. Specifically, just how close to 1.0 should lift values be for an association rule to be screened out? The recommended approach is to keep only those association rules with certain values of lift in a particular top or bottom percentage of all lift values. Similar to a scree plot of eigenvalues in the context of factor analysis, plotting a histogram of all lift values is a useful tool to help decide on these absolute or percentage values. As an illustration, Baralis et al. (2011) used such histograms and decided to further look into association rules with lift values above 10 and below 0.5.

Support is defined as P (A ∩ B) and is usually expressed in percentage metric ranging from 0 to 100. Support is the probability that A and B co-occur. Because of support's conceptual proximity to the definition of an association rule, support values should be examined before confidence values, even though both indexes are measures of an association rule's effect size. A disadvantage of support is that its usefulness decreases in the presence of very large (e.g., containing millions of transactions) and rich (e.g., containing thousands of items) data sets (E. Cohen et al., 2001). In such situations, on average, support values are likely to be quite low because the presence of other transactions (involving other items) serves as noise in the data set. In turn, because the support values of association rules will tend to be low, and given that they cannot be lower than 0, the support values of different association rules within a single data set will tend to be quite similar. Thus, the similar support values will be less useful in distinguishing the strength of one association rule from the strength of another association rule.

Confidence is defined as P (A ∩ B) / P (A) and, similar to support, is usually expressed in percentage metric ranging from 0 to 100. Confidence is the probability that a set of items is chosen given that another set of items is already chosen. An advantage of confidence is that, contrary to support, it remains sizable even in data sets that are very large and rich. Confidence avoids dependence on sample size or richness by considering only those transactions involving A and B: If A is selected, what is the probability that B is selected as well? In addition, because only the focal items are considered, that is, confidence is calculated by dividing support by P (A), an association rule's confidence value is almost always larger than its support value. As a result, even if a number of association rules do not notably differ from each other in terms of their support values, they are likely to differ in terms of their confidence values. Thus, compared to support, confidence is capable of more clearly detecting differences in the strengths of association rules. Another advantage of confidence is that it is potentially useful for building causal theories. This capability is derived from the fact that confidence can be calculated in two different ways: (a) the probability of A given B and (b) the probability of B given A. These two confidence values can in fact be drastically different, especially if the frequency of A (or B) is much larger than B (or A) (Gu, Li, He, Williams, Hawkins, & Kelman, 2003). For example, if P (A ∩ B) / P (A) is 75%, but P (A ∩ B) / P (B) is 5%, the researcher may interpret this finding as a basis for deriving the causal hypothesis that A leads to B, not the other way around (Merceron & Yacef, 2003: 205).

Similar to the interpretation of effect sizes when using more traditional approaches, the interpretation of effect sizes when using MBA is context specific (Aguinis, Werner, Abbott, Angert, Park, & Kohlhausen, 2010). A particular value for confidence may be considered important in one context but not in another context. For example, Hsieh, Lai, Lee, Hu, Tseng, and Wang (2008) defined "interesting" rules as the top five association rules on the basis of the highest confidence values. Goh and Ang (2007) designated 1% as the minimum support level and 40%, 50%, and 60% as three threshold levels for confidence values. Yang, Tang, and Kafatos (2007) used minimum cutoff values of 1.3% for support and 47.6% for confidence. As yet another example, Shahrabi and Neyestani (2009) used values of 20% and 80% for support and confidence, respectively, in defining meaningful association rules. In short, much like in the "customer-centric" approach advocated by Aguinis et al. (2010) and adopted by others (e.g., Mawritz, Mayer, Hoobler, Wayne, & Marinova, 2012; Nesterkin & Ganster, in press; Schiele, Veldman, & Hüttinger, 2011; Umphress, Bingham, & Mitchell,

Table 1
Illustrative Data Used to Compute Lift, Support, and Confidence

Transaction	Soda (A)	Orange Juice (B)	Fat-Free Cream (C)	Donuts (D)
1	1	1	1	1
2	1	1	1	1
3	1	1	1	1
…				
10	1	1	1	1
11	1	1	1	0
12	1	1	1	0
13	1	1	1	0
…				
30	1	1	1	0
31	1	0	1	0
32	1	0	1	0
33	1	0	1	0
34	0	1	1	0
35	0	1	1	0
…				
500	0	1	1	0
501	0	0	0	1
502	0	0	0	1
503	0	0	0	1
…				
990	0	0	0	1
991	0	0	0	0
…				
10,000	0	0	0	0

2010), values for confidence and support, as well as their practical significance, should be placed within a particular substantive research and application context. The same value for confidence may be interpreted as practically meaningful in a study involving the outcome "rate of survival for a particular cancer-related treatment" and not in another involving the outcome "job satisfaction."

Computing Lift, Support, and Confidence: An Illustration

In this section, we offer an illustration of the computation and interpretation of lift, support, and confidence using a hypothetical data set involving a total of 10,000 purchase transactions, a few rows of which are included in Table 1. Each row represents a single transaction, where 1 denotes *purchased*, and 0 means *not purchased*. Though there are many other items, four are considered here: soda (A), orange juice (B), fat-free cream (C), and donuts (D).

As shown in the first 30 rows in Table 1, soda (A) and orange juice (B) are purchased together in 30 of the 10,000 transactions, soda (A) is purchased in 33 transactions, and finally, orange juice (B) is purchased in 497 transactions. Accordingly, the lift of the association rule between soda (A) and orange juice (B) is $\frac{P(A \cap B)}{P(A) * P(B)} = \frac{30/10,000}{(33/10,000) * (497/10,000)} =$

18.29. Because lift is much greater than 1.0, we conclude that soda (A) and orange juice (B) are positively related.

Table 1 also shows that fat-free cream (C) and donuts (D) are purchased together in 10 of the 10,000 transactions, fat-free cream (C) is purchased in 500 transactions, and finally, donuts (D) are purchased in 500 transactions. Therefore, the lift of the association rule between fat-free cream (C) and donuts (D) is $\frac{P(A \cap B)}{P(A) * P(B)} = \frac{10/10{,}000}{(500/10{,}000) * (500/10{,}000)} =$ 0.4. Because lift is smaller than 1.0, we conclude that the two items, fat-free cream (C) and donuts (D), are negatively related (i.e., those who purchase fat-free cream are less likely to buy donuts). Given that lift values indicate that these associations exist, we next turn to obtaining support and confidence values to gain a better understanding of the strength of these relationships.

Support for the positive association between soda and orange juice is P (A ∩ B) = 30/10,000 = .3%, and support for the negative association between fat-free cream and donuts is 10/10,000 = .1%. In this example, on the basis of support values alone, it is rather difficult to make sense of any existing differences regarding these two effect sizes. On the basis of lift values, we know that one relationship is positive while the other one is negative; however, their support values seem barely distinguishable.

Confidence for the association between soda and orange juice can be calculated in the following two ways: (a) the probability of purchasing orange juice (B) given that soda (A) is also purchased = $\frac{30/10{,}000}{33/10{,}000}$ = .91 = 91%, and (b) the probability of purchasing soda (A) given that orange juice (B) is also purchased = $\frac{30/10{,}000}{497/10{,}000}$ = .06 = 6%. Because P (A∩ B) / P (A) is 91%, whereas P (A ∩ B) / P (B) is only 6%, this may serve as a basis for formulating the hypothesis that the purchase of soda causes a greater inclination to purchase orange juice. For the supermarket owner, this causal possibility suggests a potentially lucrative opportunity to boost revenue from orange juice sales by promoting soda. The reason is that the previously calculated support for this association rule is only .3%. This low figure, instead of making the association rule unimportant, may actually signal that the store is currently not being managed in a way that causes customers to buy more soda (A), which P (A ∩ B) / P (A) (= 91%) suggested might lead to greater orange juice sales. In this sense, the association rule may also constitute untapped knowledge that can help generate change in the future, as opposed to confirming what many practitioners already know (Mohrman & Lawler, 2012). Therefore, examining both the support and confidence values of the association rule between soda (A) and orange juice (B) has made it possible to gain a deeper understanding of the association rule's effect size as well as its practical significance (Aguinis et al., 2010).

Confidence values for the association rule between fat-free cream and donuts can be calculated in the following two ways: (a) the probability of purchasing donuts (D) given that fat-free cream (C) is already purchased = $\frac{10/10{,}000}{500/10{,}000}$ = .02 = 2%, and (b) the probability of purchasing fat-free cream (C) given that donuts (D) are already purchased = $\frac{10/10{,}000}{500/10{,}000}$ = .02 = 2%. Given P (C ∩ D) / P (C) = P (C ∩ D) / P (D) = 2%, and that P (C) = P (D) = 5%, we conclude that the purchase of one of the two items is associated with a decrease in the

probability of the other item's purchase by 60% = $\frac{5-2}{5}$ * 100%. Furthermore, because the two confidence values for the association rule are equal, the researcher may formulate the hypothesis that the two products repel each other to a roughly equal degree (i.e., reciprocal causality). In the alternative, the researcher may hypothesize the existence of two subgroups (e.g., health nuts vs. junk-food enthusiasts), such that group membership is conceptualized as driving the negative relationship between the two products. For the supermarket owner, this causal possibility may be interpreted as a warning not to promote the two items together in the future—possibly constituting untapped knowledge that can help prevent undesirable changes (Mohrman & Lawler, 2012).

More recently, the use of MBA has extended beyond the field of marketing. These successful applications of MBA in other scientific fields are good indicators of the potential of using MBA in management research. For example, MBA has been used to address questions of theoretical and practical importance in finance, telecommunications, and web analysis (Y.-L. Chen et al., 2005); geophysics (Yang et al., 2007); and legal aid services (Ivkovic, Yearwood, & Stranieri, 2002). Moreover, MBA has been used in cross-disciplinary research examining associations among ocean, land, and atmospheric processes (Tan, Steinbach, & Kumar, 2005). Research involving complex multilevel processes (e.g., Potter et al., 2003) also points to MBA's potential for studying similarly complex phenomena involving both micro- and macrolevel variables in management research, thereby highlighting MBA's potential to help bridge the micro–macro divide mentioned earlier.

Table 2 includes a brief summary of selected applications of MBA in a variety of scientific fields and settings. The information included in this table illustrates that the use of MBA has led to important substantive results with meaningful implications for practice in several research domains. Next, we turn to a more detailed description of the advantages of using MBA and its potential for the field of management.

Advantages of Using MBA in Management Research

MBA Allows for Inductive Theorizing

MBA is a powerful technique for inductive theorizing. More specifically, MBA allows researchers to make use of vast stores of data for theory-building purposes. The importance of the inductive capability of MBA was highlighted by Locke (2007), who argued that inductive approaches are essential for the development of theory. The impact and influence of inductive theorizing are also supported empirically by the finding that theory-building elements in an empirical article lead to a greater number of citations compared to the presence of theory-testing elements (Colquitt & Zapata-Phelan, 2007). Using MBA has the potential to lead to important contributions by allowing researchers to implement an inductive approach to theory building, which, in spite of its advantages (Locke, 2007), is currently underutilized in management research (Shepherd & Sutcliffe, 2011).

Table 2
Illustrations of Market Basket Analysis (MBA) Use in Bioinformatics, Nuclear Science, Pharmacoepidemiology, Immunology, Geophysics, and Other Fields

Source	Brief Summary of Results
Cerrito, P. B. 2007. Choice of antibiotic in open heart surgery. *Intelligent Decision Technologies*, 1: 63-69.	Surgeons prescribe antibiotics following open heart surgery, thereby reducing the incidence of postsurgery infection. Cerrito used MBA to understand which sets of antibiotics were associated with higher and lower rates of infection following heart surgery, leading to recommendations on how to improve postsurgery prescription practices.
Chen, Q., & Chen, Y.-P. P. 2006. Mining frequent patterns for AMP-activated protein kinase regulation on skeletal muscle. *BMC Bioinformatics*, 7(394): 1-14.	Chen and Chen examined possible associations between adenosine monophosphate–activated protein kinase (AMPK) and energy demand and supply in muscles. Using MBA allowed for the identification of previously unknown relationships between AMPK and metabolic actions, making AMPK a promising pharmacological target for disease treatment.
Hibino, A., & Niwa, Y. 2008. Graphical representation of nuclear incidents/accidents by associating network in nuclear technical communication. *Journal of Nuclear Science and Technology*, 45: 369-377.	Nuclear safety is of particular concern in Japan due to its population density and earthquake activity. Large amounts of data have been compiled and made available by nuclear power plants. Hibino and Niwa used MBA to map information on nuclear accidents from Japan's Nuclear Information Archives to particular website interfaces to improve how nuclear safety information is communicated to the public. MBA enabled the researchers to determine which information people found most valuable.
Hsieh, S.-C., Lai, J.-N., Lee, C.-F., Hu, F.-C., Tseng, W.-L., & Wang, J.-D. 2008. The prescribing of Chinese herbal products in Taiwan: A cross-sectional analysis of the national health insurance reimbursement database. *Pharmacoepidemiology and Drug Safety*, 17: 609-619.	Chinese herbal products are consumed widely despite the dearth of pharmacoepidemiological information on them. Unlike other drugs, herbal products are not monitored for interactions and withdrawn from the market when problems are discovered. Hsieh et al. used MBA on a data set provided by the Taiwanese National Health Insurance system that included thousands of individuals who had been prescribed Chinese herbs. The use of MBA enabled Hsieh et al. to establish coprescription patterns that require further investigation to ensure the safe consumption of Chinese herbs.
Kanagawa, Y., Matsumoto, S., Koike, S., & Imamura, T. 2009. Association analysis of food allergens. *Pediatric Allergy and Immunology*, 20: 347-352.	Individuals with food allergies are known to react to clusters of foods, resulting in some cases in the life-threatening condition of anaphylaxis. Kanagawa et al. used MBA to identify common individual food allergens as well as combinations of allergens that have identical proteins. They also ruled out some alleged allergens that were not found to be associated with each other.
Yang, R., Tang, J., & Kafatos, M. 2007. Improved associated conditions in rapid intensifications of tropical cyclones. *Geophysical Research Letters*, 34: 1-5.	Forecasting cyclones is a difficult task because of their tendency for rapid intensification (RI), which leads to high rates of error. Yang et al. used MBA on a large data set of cyclone conditions to derive hypotheses regarding which particular conditions are the best predictors of RI.

Indeed, using MBA for inductive theory building has led to insights in marketing and other fields. For example, Russell et al. (1999) pointed out that the use of MBA has allowed marketing researchers to build theoretical models of purchasing decisions involving products in more than one category (i.e., multiple-category decision making). Other researchers have also used MBA inductively to study a number of important questions outside the field of marketing. For example, Kanagawa et al. (2009) conducted research involving patients with food allergies, who are known to present allergy symptoms to multiple food allergens. The use of MBA allowed Kanagawa et al. to inductively create models regarding which allergens are related to which (e.g., chicken with eggs, abalone with salmon eggs, and matsutake mushrooms with milk).

MBA can also be used inductively to generate hypotheses that are then tested deductively in follow-up work. For example, Takeuchi, Subramaniam, Nasukawa, Roy, and Balakrishnan (2007) collected archival data from a car rental company to derive association rules that related certain types of phrases uttered by customer service agents to whether customers later picked up reserved cars. To test whether certain phrases did indeed improve the probability of pickup, Takeuchi et al. conducted a follow-up experiment to draw causal inferences. Customer service agents were randomly assigned to two groups: those who were trained on the recommendations based on the derived association rules (treatment) and those who were not (control). The dependent variable was average pickup ratio, defined as the ratio of the number of actual pickups to the number of reservations. Results were analyzed using a general linear model (GLM)–based technique (i.e., two independent-group *t* tests) and indicated that the average improvement in the pickup ratio of the treatment group was greater than that of the control group. In short, this study offers a good illustration of the possibility of first using MBA to inductively generate hypotheses and then conducting a follow-up study that is deductive in nature and uses GLM-based data-analytic procedures.

MBA Can Address Contingency Relationships

MBA can be used to uncover contingency relationships, also labeled moderating or interaction effects (Boyd et al., 2012). Specifically, MBA can reveal the strength of a given relationship as well as the extent to which such relationships (i.e., association rules) vary across different contexts. For example, suppose that using MBA yields the association rule linking variables A and B from a data set. Once a binary moderator (consisting of groups G1 and G2) is added to the data set, MBA may reveal that the confidence of [(A), (B)], or the association rule relating A and B, is weaker for G1 and stronger for G2. In addition, MBA may reveal that a distinct association, [(C), (B)], is also an association rule for G1 but not for G2 and that yet another distinct association, [(F), (A)], constitutes an association rule for G2 but not G1.

As an example, Tang, Chen, and Hu (2008) designed an algorithm that conducts MBA using a contingency approach, such that time and place were conceptualized as moderators. Adding contextual information to an MBA analysis allows for the identification of purchase patterns, taking into account whether they occur on particular days and in particular regions. Although management scholars have relied primarily on methods such as moderated multiple regression to assess contingency relationships (Aguinis, Beaty, Boik, & Pierce, 2005), MBA

is in many situations superior because it does not rely on assumptions that are often untenable, as we describe next.

MBA Does Not Rely on Often Untenable Assumptions

MBA is not bound by the strict, and often untenable, assumptions such as the linearity, normality, and residual equal variance required by the GLM and other frequently used data-analytic methods in management research (Weinzimmer, Mone, & Alwan, 1994). MBA assesses relationships between items or categories as opposed to linear relationships between two or more variables. As a result, MBA is free from the strict assumptions that are often violated in management research (Weinzimmer et al., 1994).

Clearly, as is the case for any methodological approach, MBA does have certain requirements that must be met. We describe these requirements later in the sections titled "Steps Involved in Using MBA" and "Potential Pitfalls in the Use of MBA."

MBA Allows the Use of "Unusable" and "Messy" Data

MBA enables researchers to use data that are often seen as "unusable" for management research. Most organizations regularly collect data on many business functions, such as human resources (Davenport, Harris, & Shapiro, 2010). For example, given the availability and affordability of data storage systems, organizations regularly collect data on employees (e.g., performance, absenteeism, benefits choices, training opportunites), customers (e.g., purchasing choices, frequency of visits), and many other issues. Moreover, such data are often collected unsystematically, sporadically, and without a particular scientific study in mind. MBA is ideally suited to be used inductively with such data sets to uncover association rules that may not be readily apparent (Hafley & Lewis, 1963; Shmueli, Patel, & Bruce, 2010).

"Messy" data often involve issues such as missing values and outliers. This type of data is common and poses many challenges in management research, particularly at the macro level of analysis (McDonald, Thurston, & Nelson, 2000; Roth, Switzer, & Switzer, 1999). While MBA is not immune to the problem of missing values, it allows for the interpretation of missing data as indicating that no option was selected or preferred. As previously discussed, MBA does this by deriving association rules with values of lift below 1.0, such that these association rules use the presence of one item to predict the absence of another item. Alternatively, an MBA open-source software called Weka (http://www.cs.waikato.ac.nz/~ml/weka/) contains a function called DecisionStump that treats missing values as distinct items that MBA can either predict or use to predict other items (Frank et al., 2005), thereby deriving information from both the presence and the absence of data. Thus, MBA allows researchers to derive association rules such as the following: If a customer orders a tofu vegan meal as the main dish, there is a 65% chance that the customer will not order any dessert. A caveat to this advantage is that missing responses, or nonchoices, must be substantively meaningful, as opposed to caused by artifacts, for association rules that treat the absence of an item as a distinct item to be meaningful (see the section titled "Step 4: Check MBA Requirements").

Another advantage of MBA related to messy data is that association rules are less influenced by outliers compared to more traditional data-analytic approaches. In the context of MBA, outliers result in infrequently occurring associations (He, Xu, Huang, & Deng, 2004). For example, results may indicate that only one customer's "basket" included a washing machine and a warranty for a computer. Such a combination, whether the result of an error in data entry or a customer behavior that is extremely uncommon, will have a lift value near unity and very small values for support and confidence. Consequently, this outlier will not influence overall substantive results to the extent that an outlier would influence GLM results based on standard errors, correlation coefficients, and regression coefficients (J. Cohen, Cohen, West, & Aiken, 2003).

We fully acknowledge that no method can overcome errors in data collection or entry. Nonetheless, MBA offers researchers some degree of flexibility while attempting to make full use of an extant data set. Researchers bound by GLM's assumptions will find many data sets unusable, despite the value those data sets might offer as a window into organizational phenomena that may be otherwise unavailable.

MBA Can Help Build Dynamic Theories

The field of management is increasingly acknowledging the important role of time in theory building (Mitchell & James, 2001). Accordingly, there is a growing interest in understanding phenomena as they unfold longitudinally (Ployhart & Vandenberg, 2010). Another advantage of MBA is that it can help build theories from not only cross-sectional but also longitudinal data. Thus far, we have made no distinction regarding when the transactions occur. However, MBA allows for theory-building efforts that are both cross-sectional and longitudinal (i.e., dynamic) in nature.

Given the nature of the available data, there are mainly two ways of building dynamic theories via MBA: (a) multiple MBA and (b) sequential MBA (SMBA). The multiple MBA approach is used when the available data include transactions as they have occurred over time (i.e., transactions at Time 1, transactions at Time 2, . . . , transactions at Time k). The multiple MBA approach involves first treating each data wave as a single cross-sectional study and then examining whether the lift, support, and confidence of association rules vary over time, which can be done largely descriptively by using graphs (Tang et al., 2008). SMBA is used when the available data describe individual events (i.e., items) as they have occurred over time. For example, a data set may consist of employees' files since the moment they were hired. So, the data set may include information gathered during their first, second, third, and so forth, month of employment. SMBA may uncover the presence of an association rule such as [(A), (B), (C)], suggesting that there is a pattern in which Event A (e.g., being assigned to a formal mentor) occurs before B (e.g., volunteering to participate in a particular training and development program), which occurs before C (e.g., receiving higher performance evaluation scores) (Han, Kim, & Sohn, 2009).

MBA Can Be Used to Assess Multilevel Relationships

MBA can be applied across all levels of analysis ranging from within-individual to firm-, industry-, and country-level contexts. For example, MBA can be used to simultaneously examine categories representing individual as well as group, team, or organizational characteristics. In the field of marketing, if the organizational unit is used as the level of analysis, then researchers can derive association rules linking lower-level (e.g., individual Subway restaurants) with higher-level units (e.g., entire chain of Subway restaurants within a particular state) and even higher levels of analysis (e.g., all Subway restaurants within the United States).

Methods for assessing relationships between lower-level predictors and higher-level outcomes are nascent and in early stages of development (Croon & Van Veldhoven, 2007). However, many management theories posit such effects (e.g., Coff & Kryscynski, 2011; Foss, 2011). Using MBA has great potential to lead to theoretical advancements regarding the presence of such relationships.

MBA Is Practitioner Friendly

The widely documented science–practice divide in the field of management is due, at least in part, to a lack of good communication between researchers and practitioners, who seem to speak different languages (Aguinis et al., 2010). MBA is practitioner friendly because of how results are presented. Numerical results produced by MBA are intuitive and easy to understand from a practical significance perspective. For example, support and confidence are based on a probability scale from 0% to 100%, which facilitates interpretation of association rules' practical significance (Aguinis et al., 2010). In other words, MBA relies on the strength of the association rules, which makes results more easily understandable by practitioners.

Another indication of its practitioner friendliness is that, over the course of its development, MBA has proved to be particularly useful because it offers actionable insights. Recall how immediately applicable recommendations were derived from our earlier example involving purchase transactions of soda, orange juice, fat-free cream, and donuts. When researchers are able to use MBA in their own work and collaborate with practitioners to produce insights that are relevant and actionable in terms of applications, both sides benefit, and as a result, we may see a narrowing of the science–practice divide.

In short, MBA has great potential as a methodological approach in the field of management because it allows for inductive theorizing, can address contingency relationships, does not rely on assumptions that are often violated when using GLM-based techniques, allows for the use of data often considered "unusable" and "messy," can help build dynamic theories, is suited to examine relationships across levels of analysis, and is practitioner friendly. Next, we provide a step-by-step description of how to use MBA.

Steps Involved in Using MBA

In this section, we describe the steps involved in conducting a study using MBA. The six steps that we describe next cover the entire research process, which begins with an assessment of MBA's suitability to study the issue at hand (Step 1) and the definition of the transactions (Step 2). Then, after data are collected (Step 3), there is a need to check whether some basic requirements for the use of MBA are met (Step 4). Then, association rules are derived (Step 5). Finally, results are interpreted (Step 6).

To make our description more vivid, we use a data set of 1,000 employees regarding their choices in terms of benefits. Readers can replicate our analyses and results by using the data file available at http://mypage.iu.edu/~haguinis. Although we created this data set for the particular purpose of our illustration, these data are realistic because we patterned the distribution of benefits using results from a national compensation survey (U.S. Bureau of Labor Statistics, 2006). Our illustrative data set also includes information on other issues pertaining to each of the employees (e.g., gender, number of children). Such data are usually available to human resource management units of most organizations and are of interest to researchers and practitioners wishing to study benefit bundles selected or preferred by different subgroups of employees based on demographic variables. Studying such data with MBA would be particularly helpful because a traditional one-size-fits-all approach to benefits is no longer tenable for the 21st century's increasingly diverse and global workforce (Martocchio, 2011).

Step 1: Determine Suitability of MBA

As noted earlier, MBA is typically used to examine associations based on data with binary variables. In addition, MBA can derive association rules that incorporate data collected from Likert-type or truly continuous scales also, as we described earlier (e.g., Aumann & Lindell, 2003; Bennedich, 2009). However, it would not be possible to apply MBA when the data set consists of data collected from continuous or Likert-type scales only. There must be at least some categorical variables among which there is some interest or meaningfulness for deriving categorical or quantitative association rules. Nonetheless, the majority of data routinely collected by organizations as part of their business analytics efforts (Davenport et al., 2010), as well as data collected by scholars for research purposes, contain a combination of both binary and continuous variables. Thus, MBA has potential to be used with many types of data available to organizational science researchers.

In addition, MBA has the greatest potential if the situation allows a researcher to capitalize on one or more of the advantages mentioned in the previous section. In other words, if the researcher is interested in theory building; if applying GLM-based techniques to the data set will violate strict assumptions, such as the linearity, normality, and residual equal variance; if the data seem unusable or too messy for use with traditional GLM-based approaches; if there is an interest in examining multilevel, contingency, and dynamic relationships; and if there is an interest in producing results that can be communicated easily to a practitioner audience, then MBA is likely to be a good methodological alternative. On the other hand, MBA is not appropriate for theory-testing purposes due to its exploratory nature.

Our particular illustration meets each of these criteria because we would like to inductively understand various choices employees make in terms of their benefits. Also, our illustration reflects various issues and levels of analysis that can arise when using data collected via an organization's human resource planning software such as Oracle's People Soft—data that would not typically be seen as "usable" by management researchers and data that were not collected with a particular research study in mind. Finally, we plan to report results such that if a real organization had provided the data to us, it would see the practical meaning and significance of our findings.

Step 2: Define the "Transactions"

In the field of marketing, a typical transaction consists of a set of products purchased by a customer at a retail store or on a website. These transactions constitute the observations or cases that make up the data entered into the MBA software. These items or categories can include information on individuals, teams, units, organizations, industries, or even countries. Also, transactions can take place at one point in time or over time and could involve a day, a quarter, a fiscal year, or even longer periods. Because transactions are not limited to a particular event that would be akin to stepping up to the register and checking out, transactions can be captured at any time. In other words, transactions can involve multiple levels of analysis and be part of a cross-sectional or a longitudinal data collection effort.

In the field of management, a transaction refers to a set of choices, resources, or other characteristics of a study's unit of analysis, such as an employee, entrepreneur, unit, or firm. A transaction in our specific data set refers to the set of benefits, which in a sense are choices, made available to an individual employee. Much as a supermarket customer might visit a store and fill a basket with a set of groceries (one example of a transaction), an employee selects a set of benefits when joining an organization and periodically is able to update these choices.

Step 3: Collect Data

MBA was originally designed for use with large data sets that are usually collected by others (i.e., not the research team). Thus, the researcher can seek out partnerships with organizations that will provide data in exchange for conducting analyses and presenting results to those who provided the data. It is also possible for researchers to collect their own data. However, given the effort and time involved and issues of accessibility to data sources, most MBA studies are likely to use data that have been collected by other parties (e.g., firms, chambers of commerce, professional organizations). As summarized in Table 2, researchers have been able to create data-sharing partnerships in many different fields. Furthermore, given MBA's ability to produce results that are easily understood by practitioners, we do not anticipate that establishing partnerships for the purpose of conducting management research would be difficult. Moreover, several sessions at the 2012 Academy of Management meetings discussed the need for novel methodological approaches that would allow researchers

to take advantage of available databases that are largely underutilized (Gowan et al., 2012; Siegel, Bloom, Lane, Foster, & Waldman, 2012).

In addition to using data collected by organizations, researchers can implement a third-party data collection strategy consisting of reliance on publicly available information. For example, O'Boyle and Aguinis (2012) investigated issues regarding individual performance by using data on academics, entertainers, politicians, and amateur and professional athletes. Although they did not use MBA, all of the data used in five separate studies by O'Boyle and Aguinis regarding 632,599 individuals included in 198 separate samples were gathered from publicly available websites.

Step 4: Check MBA Requirements

The fourth step is to check two key requirements of MBA (Marakas, 2003). First, it is necessary to have a sufficiently large number of transactions, because otherwise the researcher would find few, if any, association rules with values of lift meaningfully different from 1.0. Fortunately, it is fairly straightforward to meet the sample size requirement because of the vast stores of data being collected by firms in response to the analytics movement (Davenport et al., 2010) as well as the sharp decrease in cost of data storage technology (Shmueli et al., 2010). In fact, Berry and Linoff (2004) argued that firms are faced with the problem of too much data rather than too little.

Note that it may be the case that the sample size is very large, but data are collected in only one context, such as a single organization with a very unique culture. So, even a very large sample size would not mitigate concerns about the generalizability of results to other organizations (i.e., external validity evidence). In these situations, it is important for researchers to explain in detail the sources of their data so that readers can be fully informed regarding the extent to which association rules may generalize to other settings. Regarding specific sample sizes typically used in MBA studies in marketing and other fields, Berry and Linoff (2004) noted that it is not unusual for studies to include sample sizes ranging from tens of thousands to millions of transactions. However, published MBA studies have used samples as small as a few hundred transactions (e.g., Goh & Ang, 2007).

Regarding MBA's second requirement, it must be verified that nonresponses are substantively meaningful (e.g., an employee chose not to sign up for a benefit, thereby creating a nonchoice), as opposed to being caused by artifacts (e.g., highly productive and busy employees did not have the time to completely fill out a questionnaire asking about their employee benefit options, thereby creating missing responses). If it is clear that all nonresponses are substantively meaningful, then the researcher can proceed to interpret the association rules. Also, if it is possible to pinpoint which nonresponses are not substantively meaningful, then any association rule involving items with nonresponses caused by artifacts should either be interpreted with special caution or discarded.

In some situations, it may be unclear whether the nonresponses are substantively meaningful, and therefore, it will be difficult or impossible to tell from the data set or a description of the data set whether the lift, support, and confidence of association rules are inflated, deflated, or accurate. If so, it is necessary to check whether nonresponses are nonchoices

(i.e., substantively meaningful) or missing responses (i.e., caused by artifacts). To do so, the researcher can use exploratory data techniques, such as frequency tables, to identify binary-choice items to which no or few respondents responded and subsequently study the identified binary-choice items via the partnership forged with the organization that originally collected the data (Chiu & Tavella, 2008). For example, with cooperation from the organization, the researcher can examine the format of surveys used to collect the data (e.g., a binary-choice item was mistakenly excluded from the survey such that nonresponses to the item are not substantively meaningful). Also, the researcher can study the wording of items in the surveys (e.g., an item was worded in a particularly incomprehensible manner that discouraged responding). The researcher can also interview knowledgeable organizational members (e.g., a telephone conversation with a human resource manager reveals that an employee benefit option was indeed unpopular with most employees).

Our illustrative data set meets both requirements. It has a sufficiently large number of transactions because it includes 1,000 employees. Also, all nonchoices are substantively meaningful, because they denote the fact that an employee decided to sign up for one benefit option versus another. Thus, it is appropriate to move forward with the derivation and interpretation of association rules—two issues that we address next.

Step 5: Derive Association Rules and Their Strength

The researcher enters all transactions into an appropriate software program to extract association rules as well as corresponding indexes. As discussed earlier, the three most commonly used MBA indexes are lift, support, and confidence. In doing this, the researcher will likely be asked by the software program to specify cutoff values for lift, given that it is a decision-making tool used to conclude whether the presence of an association cannot be explained by chance alone. This option can be useful in preventing the program from producing an excessive number of associations (e.g., thousands or millions) from a very large and rich data set. Trying to interpret every association in such a scenario would be cumbersome. We reiterate that a histogram of all lift values is a useful tool to help decide on absolute or percentage cutoff values for lift in the given study (Baralis et al., 2011).

There are several MBA software packages commercially available, such as IBM SPSS Modeler (formerly called Clementine) and SAS Enterprise Miner. As a third commercially available option, Excel's pivot table function can also be used for a modified version of MBA (Ting, Pan, & Chou, 2010). In addition, MBA software programs that are available free of charge include arule (available online at http://www.kdnuggets.com/software/associations.html) and Magnum Opus (a demonstration version is available at http://www.giwebb.com/). Each of these software programs is accompanied by a tutorial. Because our goal in describing this fifth step is to provide an overview of the procedures involved, we will not repeat here the details that are readily available in the tutorials. Nonetheless, we strongly encourage readers to use the tutorials and data sets made available by the software packages because, as is the case with any new methodological approach, it will be difficult to fully understand how to use MBA and interpret the resulting software output without actually trying it.

In our experience, all of these programs have similar capabilities, although they vary in terms of the particular labels used for certain procedures and results. Because of the similarities across software programs in terms of their overall capabilities, the choice of one package over another is mostly an issue of personal preference and familiarity with each package. So, for example, the learning curve for IBM SPSS Modeler will be less steep for IBM SPSS users. Thus, our recommendation is to use a software package with which a researcher is already familiar.

Step 6: Interpret Association Rules

The next and final step is to interpret the derived association rules. In our particular example, we used Magnum Opus. Using our data set, and for the pedagogical purpose of the illustration, we constrained our analysis to rules involving pairs of items only and positive associations only (i.e., as discussed earlier, as indicated by lift values larger than 1.0). Results indicated a total of 28 association rules, and we sorted them by lift value from highest to lowest.

From these results, the three rules with the highest lift values are (a) [(Dependent Care Reimbursement), (Wellness Programs)], with lift of 2.42, support of 4%, and confidence (called "strength" by Magnum Opus), or the probability that the right-hand side item is chosen given that the left-hand side item is chosen, of 30.8%; (b) [(Holiday Pay), (Retirement Plans)], with lift of 1.94, support of 11.1%, and confidence of 42.2%; and (c) [(Holiday Pay), (Health Care Reimbursement)], with lift of 1.85, support of 8.5%, and confidence of 32.3%.

Because all three association rules have lift values much higher than 1.0, we can conclude that these rules exist in the form of positive relationships between the items and are not due to chance. We can then consider the support values for these three rules, which range from 4% to 11.1%. Recall that support is the proportion of transactions in which two items appear together, so that the items in these rules co-occur in 4% to 11.1% of the transactions in this data set. [(Holiday Pay), (Retirement Plans)] and [(Holiday Pay), (Health Care Reimbursement)] have the highest support values out of the 28 association rules returned by our criteria, whereas support for the association rule [(Dependent Care Reimbursement), (Wellness Programs)] is in the middle of the range of all support values (ranging from 1.8% to 11.1%). The confidence values for these three rules range from 30.8% to 42.2%, given that all confidence values (i.e., probability that the right-hand side item is chosen given that the left-hand side item is chosen) in the data set range from 13.8% to 51.2%.

Using these results for theory development purposes involves consideration of causal hypotheses, which is accomplished by examining the two confidence values of each rule (i.e., the presence of Item B given the presence of Item A, as well as the presence of Item A given the presence of Item B; Gu et al., 2003). [(Dependent Care Reimbursement; A), (Wellness Programs; B)] has confidence values of 30.8% (probability of B given A) and 31.5% (probability of A given B). [(Holiday Pay; A), (Retirement Plans; B)] has confidence values of 42.2% (probability of B given A) and 51.2% (probability of A given B). Finally, [(Holiday Pay; A), (Health Care Reimbursement; B)] has confidence values of 32.3% (probability of B given A) and 48.6% (probability of A given B). Only the last example has widely

divergent confidence values, such that we may wish to investigate whether the choice to have health care reimbursement leads to the choice to have holiday pay (rather than the other way around). One implication of this finding might be that using health care reimbursement makes employees pay more attention to and subsequently become more critical about their financial status such that they begin to seek additional forms of income, such as holiday pay.

We may also infer from the two rules with the highest support—[(Holiday Pay), (Retirement Plans)] and [(Holiday Pay), (Health Care Reimbursement)]—that holiday pay is selected by employees who are seeking to maximize their income from the firm rather than focusing on work–life balance. Specifically, these employees may be willing to forgo their holiday time for money because of financial pressures, which are leading them to also seek security during retirement and protection from health care costs. These employees stand in contrast to others who do not select holiday pay and instead prefer to spend holidays with their families. In this sense, variables such as financial hardship may serve as a potential moderator.

Furthermore, having dependents may be a moderator of some of the derived associations, including the relationship between dependent care reimbursement and wellness programs. For example, the firm may discover that employees without dependents tend to choose dependent care reimbursement and wellness programs much less frequently together (e.g., support of 1%) than do employees with dependents (e.g., support of 12%). Findings of contingency relationships from MBA can thus help form contingency theories as well as competing theories. Follow-up studies pitting such competing theories against one another can further advance the employee benefits, compensation, motivation, and decision-making literatures.

Such findings from MBA constitute actionable knowledge for practitioners as well, not just knowledge for theory building by researchers, thereby narrowing the science–practice divide. Recall the previous example where the two association rules—[(Holiday Pay), (Retirement Plans)] and [(Holiday Pay), (Health Care Reimbursement)]—applied more strongly to employees under financial pressure. Practitioners at the firm where the data were collected can use the finding to take a number of actions. For instance, if a strategic objective of the firm is to provide support to employees in areas outside of their work lives, practitioners may wish to promote a financial education program to struggling employees opting for holiday pay. They may also wish to encourage social outings among these employees, whose financial strain and forgone holidays may lead to burnout that could be ameliorated with greater social engagement with their peers. Finally, because both academia and practice benefit from results derived from MBA, the method helps narrow the science–practice gap also by fostering collaborations between researchers and practitioners.

Next, we provide examples of substantive research in various management domains that can benefit from the use of MBA. Then, we discuss potential pitfalls to avoid when using MBA.

Illustrations of Management Research Domains That Would Benefit From Using MBA

In this section, we discuss three selected areas of management research that would benefit from the use of MBA. We provide one illustration in each of the following management

subfields: organizational behavior, entrepreneurship, and strategic management. Note that our previous section included an illustration regarding a substantive domain in human resource management (i.e., employee benefits). Our goal in this section is to illustrate the applicability and potential of MBA across management domains. For each illustration, we highlight how substantive research in each of these domains would benefit from MBA's advantages.

Organizational Behavior Illustration: Dysfunctional Workplace Behaviors

Organizational behavior researchers have called for studies on how a wide range of dysfunctional workplace behaviors appear, aggravate, or weaken over time. This body of research addresses the overall need to understand how dysfunctional workplace behaviors are associated with consequences for various stakeholders over time (Robinson, 2008).

MBA's fifth advantage (i.e., MBA can help build dynamic theories) can be useful in developing longitudinal models of dysfunctional workplace behaviors and their correlates. MBA is particularly suited for doing so, because firms usually possess employee records, including information on several forms of dysfunctional behavior, such as aggression, unexcused absenteeism, and incivility. Human resource departments also keep information on written warnings, demotions, and pay deductions (e.g., Dietz, Robinson, Folger, Baron, & Schulz, 2003). These data may often be seen as messy and unusable in terms of more traditional methodological approaches, but they can be used with MBA (thereby capitalizing on MBA's fourth advantage, its ability to use data typically considered "unusable" and "messy"). Specifically, this particular research study would involve using SMBA, which would allow for the identification of sequential association rules, such as [(Abusive Language), (Physical Threat)]. This may lead to the inductive development of the hypothesis that dysfunctional workplace behaviors start small and then escalate in intensity. Such sequential association rules can be used toward the development of a dynamic theory of dysfunctional workplace behaviors.

Entrepreneurship Illustration: Entrepreneurs' Identities

A promising research topic in the entrepreneur identity literature is the study of entrepreneurs' micro-identities (e.g., entrepreneur, parent, humanitarian, and friend). Because the study of micro-identities is in its inception, future research that examines basic questions, such as how entrepreneurs acquire such identities, will be able to make valuable contributions to the field of entrepreneurship (Shepherd & Haynie, 2009).

MBA is a method well suited for making such contributions. Specifically, there are vast repositories of data regarding entrepreneurs and their activities, including the Global Entrepreneurship Monitor, the Panel Study of Entrepreneurial Dynamics, and the U.S. Census Bureau (e.g., self-employment data). MBA can be used to inductively build theories on how entrepreneurs acquire and manage new micro-identities, thereby capitalizing on the first advantage of MBA (i.e., MBA allows for inductive theorizing). For instance, results of

such analyses may reveal that many dependent-free entrepreneurs tend to maintain micro-identities as both an entrepreneur and a provider of a family: [(Entrepreneur), (Provider), (No Dependents)]. Instead of rejecting this association rule that seems puzzling, such a result may lead to inductively derived propositions that can be tested deductively in follow-up research. Another association rule derived from the data could be that [(Entrepreneur), (Soldier), (No Prior Military Experience), (Family Member in Military)], may lead the researcher to construct another theory that entrepreneurs also acquire new micro-identities by adopting or mimicking the micro-identities of close others. By inductively deriving hypotheses and new theories about how entrepreneurs acquire new micro-identities, MBA can make valuable contributions to the field of entrepreneurship.

Strategic Management Illustration: Corporate Social Responsibility

Research in strategic management now recognizes the need to examine individual-level processes (i.e., microfoundations) as mediating mechanisms of traditional macrolevel relationships in order to shed light on the "black boxes" underlying macrolevel relationships (e.g., Coff & Kryscynski, 2011; Foss, 2011). One such macrolevel relationship whose microfoundation needs greater development is the corporate social responsibility–firm outcomes relationship (Aguinis & Glavas, 2012). For example, after a firm creates a policy and budget for corporate social responsibility (CSR) activities, what types of attributes lead employees to then actually initiate or volunteer for CSR projects? Questions such as this one necessitate the investigation of multilevel relationships, because CSR takes place at the organizational level of analysis whereas individual factors reside at the individual level of analysis (Aguinis & Glavas, 2012).

This type of research can capitalize on MBA's sixth advantage (i.e., MBA can be used to assess multilevel relationships). Specifically, data sources to be used for MBA-based research can include a firm's personnel records for the individual-level data. For example, JPMorgan Chase supports "more than 1,800 employee-led volunteer projects" annually (JPMorgan Chase & Co., 2012). In addition, there are several sources of publicly available databases that can be used to collect organizational-level data, such as the United Nations Global Compact (http://www.unglobalcompact.org).

Potential Pitfalls in the Use of MBA

Although we believe that the use of MBA in management research is likely to lead to important theoretical advancements and subsequent applications, we readily acknowledge that, as is the case with any methodological tool, MBA is not necessarily the best alternative for all research domains, data structures, and situations. In this section, we describe three important pitfalls that researchers should avoid when using MBA.

First, data must be collected or retained for all possible types of outcomes of interest. For example, Berry and Linoff (2004) encountered a company evaluating marketing methods that had retained information only on customers who responded to direct mail offers. The

organization had discarded data for individuals who had not responded (i.e., those who had responded to other types of offers), thereby making it impossible to assess any presence of association rules between customer characteristics and type of response medium. Note that this is not an issue of missing data for some of the respondents for a particular category but a situation involving the complete absence of responses in a category. Referring back to our illustration using benefits, a researcher would not be able to assess the possible presence of an association rule involving, for example, a particular employee benefit if data on the benefit were never collected.

Second, archival data collected by firms, professional associations, and other organizations are usually different from data collected for academic research. Specifically, third-party data sets may have errors in labeling of fields and recording of information (Berry & Linoff, 2004). Moreover, the construct validity of many of the measures may be suspect—as is often the case with archival data in general (Boyd, Gove, & Hitt, 2005). Thus, the fact that a firm provides a data file with a variable labeled "employee motivation" does not mean that the data are actually about employee motivation. In short, prior to using MBA, it is important for researchers to check the integrity of the data and the validity of the measures used to collect those data, particularly when the data have been collected by a third party and for purposes other than the particular theory-based goals sought by the researchers.

Finally, learning a new methodological tool or statistical technique requires hands-on experience. Though we provided a description of how to use MBA through an illustrative data set, it will be difficult to fully understand how to use the method without actually trying it. Similar to any other innovative methodological approach, implementation is a key issue. We encourage readers to use MBA with actual data—we believe that only then will the full extent of MBA's benefits become clear.

Conclusion

MBA is a methodological approach that originated in the field of marketing and has more recently been used effectively in fields such as bioinformatics, nuclear science, pharmacoepidemiology, immunology, and geophysics. One reason for MBA's increasing adoption across scientific fields is that it allows researchers to assess the presence of association rules by using an inductive approach to theorizing. Moreover, MBA allows researchers to address contingency relationships, does not rely on often untenable assumptions (as many other data-analytic approaches do), allows for the use of data often considered "unusable" or "messy," can help build dynamic theories, can be used to assess multilevel relationships, and is practitioner friendly. Thus, MBA has great potential in terms of producing theoretical insights in management that are also likely to lead to meaningful practices and organizational interventions. There are indicators of MBA's great potential originating from both the academic and practitioner domains. From the academic side, as mentioned earlier, the 2012 Academy of Management meetings included several sessions with the common goal of discussing the need for novel techniques that can be used to take advantage of large databases, such as the World Management Survey, STAR Metrics (Science and Technology for America's Reinvestment: Measuring the Effect of Research on Innovation, Competitiveness and

Science), and the U.S. Census Bureau's new survey of Management and Organizational Practices (MOPS; Gowan et al., 2012; Siegel et al., 2012). From the practitioner side, a common topic of conversation is that "we're entering a new era, and the change that's driving it is the rise of big data" (IBM, 2012: 11).

Due to the decrease in the cost associated with data storage, organizations routinely collect data at the individual, team, unit, and organizational levels of analysis. MBA can help narrow the science–practice divide by allowing firms to analyze data they already possess and researchers to analyze data they did not have to collect. Frequently, it is the inaccessibility of large data sets or the prohibitively high cost of collecting data that prevents advancements in scholarly research. MBA is particularly suited to tackle important research questions using those data across management subfields, including human resource management, organizational behavior, entrepreneurship, and strategic management, among others. We hope our article will serve as a catalyst for the adoption of MBA as a novel methodological approach in management research.

References

Agrawal, R., Imieliński, T., & Swami, A. 1993. Mining association rules between sets of items in large databases. In P. Buneman, & S. Jajodia (Eds.), *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*: 207-216. New York: Association for Computing Machinery.

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. 2005. Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology*, 90: 94-107.

Aguinis, H., Boyd, B. K., Pierce, C. A., & Short, J. C. 2011. Walking new avenues in management research methods and theories: Bridging micro and macro domains. *Journal of Management*, 37: 395-403.

Aguinis, H., Dalton, D. R., Bosco, F. A., Pierce, C. A., & Dalton, C. M. 2011. Meta-analytic choices and judgment calls: Implications for theory building and testing, obtained effect sizes, and scholarly impact. *Journal of Management*, 37: 5-38.

Aguinis, H., & Glavas, A. 2012. What we know and don't know about corporate social responsibility: A review and research agenda. *Journal of Management*, 38: 932-968.

Aguinis, H., Werner, S., Abbott, J. L., Angert, C., Park, J. H., & Kohlhausen, D. 2010. Customer-centric science: Reporting significant research results with rigor, relevance, and practical impact in mind. *Organizational Research Methods*, 13: 515-539.

Alvarez, S. A. 2003. *Chi-squared computation for association rules: Preliminary results*. Technical report BC-CS-2003-01. Chestnut Hill, MA: Boston College.

Aumann, Y., & Lindell, Y. 2003. A statistical theory for quantitative association rules. *Journal of Intelligent Information Systems*, 20: 255-283.

Baralis, E., Cagliero, L., Cerquitelli, T., Garza, P., & Marchetti, M. 2011. CAS-MINE: Providing personalized services in context-aware applications by means of generalized rules. *Knowledge and Information Systems*, 28: 283-310.

Bennedich, M. 2009. *Mining survey data*. Unpublished master's thesis in computer science, School of Engineering and Business Management, Royal Institute of Technology, Stockholm, Sweden.

Berry, M. J. A., & Linoff, G. S. 2004. *Data mining techniques for marketing, sales, and customer relationship management* (2nd ed.). Indianapolis, IN: Wiley.

Bowen, H. P. 2012. Testing moderating hypotheses in limited dependent variable and other nonlinear models: Secondary versus total interactions. *Journal of Management*, 38: 860-889.

Boyd, B. K., Gove, S., & Hitt, M. A. 2005. Construct measurement in strategic management research: Illusion or reality? *Strategic Management Journal*, 26: 239-257.

Boyd, B. K., Haynes, K. T., Hitt, M. A., Bergh, D. D., & Ketchen, D. J., Jr. 2012. Contingency hypotheses in strategic management: Use, disuse, or misuse? *Journal of Management*, 38: 278-313.

Cascio, W. F., & Aguinis, H. 2008. Research in industrial and organizational psychology from 1963 to 2007: Changes, choices, and trends. *Journal of Applied Psychology*, 93: 1062-1081.

Cerrito, P. B. 2007. Choice of antibiotic in open heart surgery. *Intelligent Decision Technologies*, 1: 63-69.

Chen, Q., & Chen, Y.-P. 2006. Mining frequent patterns for AMP-activated protein kinase regulation on skeletal muscle. *BMC Bioinformatics*, 7(394): 1-14.

Chen, Y.-L., Tang, K., Shen, R.-J., & Hu, Y.-H. 2005. Market basket analysis in a multiple store environment. *Decision Support Systems*, 40: 339-354.

Chiu, S., & Tavella, D. 2008. *Data mining and market intelligence for optimal marketing returns*. Oxford, UK: Butterworth-Heinemann.

Coff, R., & Kryscynski, D. 2011. Drilling for micro-foundations of human capital–based competitive advantages. *Journal of Management*, 37: 1429-1443.

Cohen, E., Datar, M., Fujiwara, S., Gionis, A., Indyk, P., Motwani, R., Ullman, J. D., & Yang, C. 2001. Finding interesting associations without support pruning. *IEEE Transactions on Knowledge and Data Engineering*, 13: 64-78.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. 2003. *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). New York: Routledge.

Colquitt, J. A., & Zapata-Phelan, C. P. 2007. Trends in theory building and theory testing: A five-decade study of the *Academy of Management Journal*. *Academy of Management Journal*, 50: 1281-1303.

Croon, M. A., & Van Veldhoven, M. J. P. M. 2007. Predicting group-level outcome variables from variables measured at the individual level: A latent variable multilevel model. *Psychological Methods*, 12: 45-57.

Davenport, T. H., Harris, J., & Shapiro, J. 2010. Competing on talent analytics. *Harvard Business Review*, 88(10): 52-58.

Dietz, J., Robinson, S. L., Folger, R., Baron, R. A., & Schulz, M. 2003. The impact of community violence and an organization's procedural justice climate on workplace aggression. *Academy of Management Journal*, 46: 317-326.

Foss, N. J. 2011. Why micro-foundations for resource-based theory are needed and what they may look like. *Journal of Management*, 37: 1413-1428.

Frank, E., Hall, M., Holmes, G., Kirby, R., Pfahringer, B., Witten, I. H., & Trigg, L. 2005. Weka. *Data Mining and Knowledge Discovery Handbook*, 8: 1305-1314.

Goh, D. H., & Ang, R. P. 2007. An introduction to association rule mining: An application in counseling and help-seeking behavior of adolescents. *Behavior Research Methods*, 39: 259-266.

Gowan, M. A., Klimoski, R. J., Schmit, M. J., Cortina, J. M., Fried, Y., Mullins, F., & Zagorsky, J. 2012. *Using large-scale archival datasets for human resource management research*. Professional development workshop conducted at the meetings of the Academy of Management, Boston.

Gu, L., Li, J., He, H., Williams, G., Hawkins, S., & Kelman, C. 2003. Association rule discovery with unbalanced class distributions. *Lecture Notes in Computer Science*, 2903: 221-232.

Hafley, W. L., & Lewis, J. S. 1963. Statistical approaches to experimental data: Analyzing messy data. *Industrial and Engineering Chemistry*, 55: 37-39.

Han, H. K., Kim, H. S., & Sohn, S. Y. 2009. Sequential association rules for forecasting failure patterns of aircrafts in Korean airforce. *Expert Systems With Applications*, 36: 1129-1133.

He, Z., Xu, X., Huang, J. Z., & Deng, S. 2004. Mining class outliers: Concepts, algorithms and applications in CRM. *Expert Systems With Applications*, 27: 681-697.

Hibino, A., & Niwa, Y. 2008. Graphical representation of nuclear incidents/accidents by associating network in nuclear technical communication. *Journal of Nuclear Science and Technology*, 45: 369-377.

Hsieh, S.-C., Lai, J.-N., Lee, C.-F., Hu, F.-C., Tseng, W.-L., & Wang, J.-D. 2008. The prescribing of Chinese herbal products in Taiwan: A cross-sectional analysis of the national health insurance reimbursement database. *Pharmacoepidemiology and Drug Safety*, 17: 609-619.

IBM. 2012. A smarter planet is built on smarter analytics. *Fortune*, 165(7): 11.

Ivkovic, S., Yearwood, J., & Stranieri, A. 2002. Discovering interesting association rules from legal databases. *Information and Communications Technology Law*, 11: 35-47.

JPMorgan Chase & Co. 2012. *Responsibility*. http://careers.jpmorgan.com/student/jpmorgan/careers/workinghere/responsibility. Accessed March 28, 2012.

Kanagawa, Y., Matsumoto, S., Koike, S., & Imamura, T. 2009. Association analysis of food allergens. *Pediatric Allergy and Immunology*, 20: 347-352.

Larose, D. T. 2005. *Discovering knowledge in data: An introduction to data mining*. Hoboken, NJ: Wiley-Interscience.

Locke, E. A. 2007. The case for inductive theory building. *Journal of Management*, 33: 867-890.

Marakas, G. M. 2003. *Modern data warehousing, mining, and visualization: Core concepts*. Upper Saddle River, NJ: Prentice Hall.

Martocchio, J. J. 2011. Strategic reward and compensation plans. In S. Zedeck (Ed.), *APA handbook of industrial and organizational psychology, vol. 1: Building and developing the organization*: 343-372. Washington, DC: APA.

Mawritz, M. B., Mayer, D. M., Hoobler, J. M., Wayne, S. J., & Marinova, S.V. 2012. A trickle-down model of abusive supervision. *Personnel Psychology*, 65: 325-357.

McDonald, R. A., Thurston, P. W., & Nelson, M. R. 2000. A Monte Carlo study of missing item methods. *Organizational Research Methods*, 3: 71-92.

Merceron, A., & Yacef, K. 2003. A web-based tutoring tool with mining facilities to improve learning and teaching. In U. Hoppe, F. Verdejo, & J. Kay (Eds.), *Artificial intelligence in education: Shaping the future of learning through intelligent technologies*: 201-208. Amsterdam: IOS Press.

Mitchell, T. R., & James, L. R. 2001. Building better theory: Time and the specification of when things happen. *Academy of Management Review*, 26: 530-547.

Mohrman, S. A., & Lawler, E. E., III. 2012. Generating knowledge that drives change. *Academy of Management Perspectives*, 26: 41-51.

Nesterkin, D. A., & Ganster, D. C. in press. The effects of nonresponse rates on group-level correlations. *Journal of Management*. doi:10.1177/0149206311433853

Nisbet, R., Elder, J. F., & Miner, G. 2009. *Handbook of statistical analysis and data mining applications*. Boston: Academic Press/Elsevier.

O'Boyle, E. H., Jr., & Aguinis, H. 2012. The best and the rest: Revisiting the norm of normality of individual performance. *Personnel Psychology*, 65: 79-119.

Ployhart, R. E., & Vandenberg, R. J. 2010. Longitudinal research: The theory, design, and analysis of change. *Journal of Management*, 36: 94-120.

Potter, C., Klooster, S., Steinbach, M., Tan, P., Kumar, V., Shekhar, S., Nemani, R., & Myneni, R. 2003. Global teleconnections of climate to terrestrial carbon flux. *Journal of Geophysical Research*, 108(D17): 1-12.

Robinson, S. 2008. Dysfunctional workplace behavior. In J. Baling, & C. Cooper (Eds.) *The Sage handbook of organizational behavior, vol. 1: Micro approaches*: 141-159. Los Angeles: Sage.

Roth, P. L., Switzer, F. S., III, & Switzer, D. M. 1999. Missing data in multiple item scales: A Monte Carlo analysis of missing data techniques. *Organizational Research Methods*, 2: 211-232.

Russell, G. J., & Petersen, A. 2000. Analysis of cross category dependence in market basket selection. *Journal of Retailing*, 76: 367-392.

Russell, G. J., Ratneshwar, S., Shocker, A. D., Bell, D., Bodapati, A., Degeratu, A., Hildebrandt, L., Kim, N., Ramaswami, S., & Shankar, V. H. 1999. Multiple-category decision-making: Review and synthesis. *Marketing Letters*, 10: 319-332.

Schiele, H., Veldman, J., & Hüttinger, L. 2011. Supplier innovativeness and supplier pricing: The role of preferred customer status. *International Journal of Innovation Management*, 15: 1-27.

Shahrabi, J., & Neyestani, R. S. 2009. Discovering Iranians' shopping culture by considering virtual items using data mining techniques. *Journal of Applied Sciences*, 9: 2351-2361.

Shepherd, D. A., & Haynie, J. M. 2009. Family business, identity conflict, and an expedited entrepreneurial process: A process of resolving identity conflict. *Entrepreneurship Theory and Practice*, 33: 1245-1264.

Shepherd, D. A., & Sutcliffe, K. M. 2011. Inductive top-down theorizing: A source of new theories of organization. *Academy of Management Review*, 36: 361-380.

Shmueli, G., Patel, N. R., & Bruce, P. C. 2010. *Data mining for business intelligence: Concepts, techniques, and applications in Microsoft Office Excel with XLMiner*. Hoboken, NJ: Wiley.

Siegel, D., Bloom, N., Lane, J., Foster, L., & Waldman, D. A. 2012. *Management practices and data sets*. Professional development workshop conducted at the meetings of the Academy of Management, Boston.

Takeuchi, H., Subramaniam, L. V., Nasukawa, T., Roy, S., & Balakrishnan, S. 2007. *A conversation-mining system for gathering insights to improve agent productivity*. Paper presented at the IEEE International Conference, Tokyo, Japan.

Tan, P.-N., Steinbach, M., & Kumar, V. 2005. *Introduction to data mining*. Boston: Pearson Education.

Tang, K., Chen, Y.-L., & Hu, H.-W. 2008. Context-based market basket analysis in a multiple store environment. *Decision Support Systems*, 45: 150-163.

Ting, P.-H., Pan, S., & Chou, S.-S. 2010. Finding ideal menu items assortments: An empirical application of market basket analysis. *Cornell Hospitality Quarterly*, 51: 492-501.

Umphress, E. E., Bingham, J. B., & Mitchell, M. S. 2010. Unethical behavior in the name of the company: The moderating effect of organizational identification and positive reciprocity beliefs on unethical pro-organizational behavior. *Journal of Applied Psychology*, 95: 769-780.

U.S. Bureau of Labor Statistics. 2006. *National compensation survey*. Washington, DC: U.S. Government Printing Office.

Webb, G. I. 1999. *Magnum Opus application help*. Glen Iris, Australia: Author.

Weinzimmer, L. G., Mone, M. A., & Alwan, L. C. 1994. An examination of perceptions and usage of regression diagnostics in organization studies. *Journal of Management*, 20: 179-192.

Yang, R., Tang, J., & Kafatos, M. 2007. Improved associated conditions in rapid intensifications of tropical cyclones. *Geophysical Research Letters*, 34: 1-5.

Zhang, C., & Zhang, S. 2002. *Association rule mining: Models and algorithms*. Berlin, Germany: Springer.