Reasons for Enthusiasm and Caution Regarding Big Data in Applied Selection Research

John D. Morrison, Jr.
Independent Consultant

Joseph D. Abraham
PSI Services LLC

Big Data and Big Data Analytics (BD and BDA, respectively) have burst onto the scene in the popular culture in recent years, perhaps most notably in the form of Nate Silver’s high profile predictions in the political and professional sports arenas. According to futurist Tyler Cowen, modern society will soon be reordered, with creators and tenders of incredibly complex datasets and the artificial intelligences mining those datasets at the top and the remaining humans relegated to a vast underclass (Cowen, 2013). SIOP’s interest in BD and BDA’s potential talent management applications has increased as well. Case in point, this article is the third on the topic to be published in TIP in the last five issues. Whereas the previous articles (Maurath, 2014; Poeppelman, Blacksmith, & Yang, 2013) viewed BD and BDA through a wide-angle lens, we focus more narrowly on their applied selection research applications. Although we see the potential for BD and BDA to provide considerable incremental value, the current article can be considered a cautionary editorial on their merits in this context. We focus particularly on assessment design and validation, highlighting areas where BD and BDA can contribute (and are already doing so) while also identifying areas where they are unlikely to bear fruit, may pose risks, or are potentially unnecessary.

The article is organized in terms of recent definitions of Big Data (Laney, 2001; Maurath, 2014). Specifically, Big Data is defined in terms of “three Vs” — volume, velocity, and variety—and we discuss each of these aspects in turn. In preparing for this article, we spoke with A. James Illingworth and Michael Lippstreu, two of the authors of the selection and assessment chapter in the forthcoming SIOP Organizational Frontiers Series book Big Data at work: The Data Science Revolution and Organizational Psychology (Illingworth, Lippstreu, & Deprez-Sims, 2015). At appropriate points throughout, we share their perspectives on Big Data’s role in applied selection research and talent acquisition/management.

Volume

According to the 3Vs definition, for data to meet this condition for “bigness” it must be so vast that it cannot be stored on a single computer’s hard drive or manipulated using typical statistical software packages (Maurath, 2014). In applied selection research, the most likely reason for a dataset to become that large is sample size. Lippstreu and Illingworth, among others, have correctly pointed out that the talent acquisition systems of large employers can easily accumulate millions of new applicant
records per year at the beginning stages of a multihurdle applicant screening process.

When it comes to big volume due to large sample sizes, the amount of data this definition describes is unnecessary for almost any purpose related to assessment design and validation. Modern computer hard drives can accommodate very large data sets. As of October 2014, $1,200 USD would buy a laptop with 16 gigabytes of RAM and a 1 terabyte internal hard drive—easily enough processing speed and storage capacity to handle datasets with tens or even hundreds of thousands of cases containing a handful of relatively simple variables. But the more important issue is related to sample sizes necessary to detect meaningful relationships between predictors and criteria. Cohen (1988) has proposed what are likely the best known rules of thumb for the interpretation of effect size magnitude in the social sciences, with suggested classifications of Pearson $r$ values as small (.10), medium (.30), and large (.50). The Employment and Training Administration of the Department of Labor’s Testing and Assessment: An Employer’s Guide to Good Practices (2000) guidelines for the likely usefulness of selection instruments are generally aligned with Cohen’s rules of thumb. A sample consisting 1,300 participants is sufficient to detect a correlation with an absolute value of .10 (Cohen’s example of a “small” effect) when alpha = .05 (two tailed) and power = .95 (as computed using standard power analysis techniques). This is a large sample certainly but not “big” per the Big Data volume definition.

Of course, larger or smaller samples may be appropriate for applied selection research, depending on the study’s objective and its practical constraints. Beyond examining test validity, researchers are sometimes interested in investigating topics requiring large samples, such as measurement invariance, measurement bias, or differential validity (e.g., Berry, Cullen, & Meyer, 2014; Meade, 2008; Meade, Johnson, & Braddy, 2008; Roth, Bobko, & Switzer, 2006; Roth et al., 2014). Yet even in cases where large amounts of data can be persuasive, existing sources and analytical methods could be sufficient and arguably preferable. It is worth keeping in mind, as Poeppelman et al. (2013) mentioned, that working with large amounts of data is not new to I-O psychology. Over the last quarter century, for example, I-O practitioners have been utilizing large aggregates of data in the form of meta-analyses to evaluate the extent to which a predictor’s validity generalizes for multiple jobs or job families across different settings. The Principles for the Validation and Use of Personnel Selection Procedures (SIOP, 2003) endorses their use for this purpose. Thus, in terms of volume, BD’s contribution to these areas of applied selection research seems to be incremental at best and likely to reach an asymptote well before challenging a modern computer’s storage or processing capacity.

**Velocity**

Another key attribute of Big Data is that it accumulates very rapidly. In the contexts of talent acquisition analytics and applied selection research, it is worth distinguishing between what we have termed between-subjects velocity and within-subjects velocity. Between-subjects velocity is the rapid collection of measurements obtained once
on many individuals. Illingworth et al. (2015) talk about the value of velocity in talent acquisition systems in their upcoming chapter on Big Data selection and assessment, focusing on what we refer to as between-subjects velocity. Specifically, at the beginning stages of a large-scale applicant screening process, many thousands of new job seekers across multiple geographic areas may view a realistic job preview video or start a job application each day. This kind of between-subjects velocity decelerates as the qualified applicant pool is winnowed at later stages. Still, within the realm of applied selection, useful data on the amount of time people are spending reviewing each assessment question, for example, will accumulate quite rapidly. Because the volume levels needed are relatively modest compared to the system’s capabilities, selection practitioners are often able to evaluate item and test-level psychometric properties of assessments very soon after implementation and monitor for changes over time. Such analyses are useful for test security purposes, as well as for ongoing monitoring of system effectiveness.

Within-subjects velocity is the rapid collection of multiple measurements on the same variable(s) on a single individual over time. In other fields where BD is utilized, several kinds of measurements are taken within fractions of a second. Maurath (2014) used the example of Google’s self-driving car, in which repeated measurements on the same variables accrue at 750 megabytes per second in order for it to stay on the road. But this is rarely if ever necessary in the vast majority of applied selection settings, where the goal is to assess stable, job-related characteristics (predictors) expected to change very little over short time periods. When we can be confident in the reliability of our tools, retesting trajectory information from within-subjects study designs can provide useful information on how response distortion/test taking strategies evolve across administrations (Schleicher, Van Iddekinge, Morgeson, & Campion, 2010). The within-subjects administration intervals in retest trajectory studies typically range from same day to a number of months. So although talent acquisition systems make it easier to collect the information needed for studies evaluating a tool’s operational performance, the required intervals fall short of Big Data velocity standards. Retesting studies involve readministering the entire assessment over time. Alternatively, within-subjects high velocity data acquisition can also take the form of repeated measures obtained on an individual during a single assessment administration. This kind of high-velocity data collection is likely to be useful in simulation assessments, particularly for jobs with a heavy psychomotor component (e.g. tracking real-time rudder adjustments made by pilots during flight simulations). But, at least for now, its usefulness in selection assessments is probably limited to such narrow applications.

The same holds for criterion measurement. Many of us have trip computers as standard features in our cars. Set accordingly, they can present a continuous real-time indicator of the car’s fuel efficiency. The values fluctuate wildly, from unnervingly low single-digit miles-per-gallon readings when accelerating from a stop or maintaining speed on an incline, to impressive, overly comforting double digits after
reaching high gear cruising speed on level roads. But when evaluating a car’s overall fuel efficiency, this information is inferior to the summary-level values provided on the sticker or the information a driver can obtain by simply dividing distance driven by fuel usage. Likewise, when practitioners need to evaluate the job performance of commercial truck drivers or pilots for use as criteria in validation research, more global indicators will typically be preferred over minute bits of data collected at infinitesimal time intervals. There are some exceptions in which performance data are collected at much tighter intervals in specific industries (e.g. call centers, manufacturing), and the practice may become more prevalent with advances in monitoring technology. However, at present, the most psychologically meaningful, valid job performance criteria will usually be more global in nature than the kinds of measures that would put one in Big Data territory.

Variety

The third element of the 3Vs definition refers to the tremendous numbers and diversity of variables available for evaluating individuals as a result of their activity in social media, elsewhere on the Internet, and from other large data sources. For the most part, this element’s benefit is to applied selection research centers on postimplementation consulting opportunities. In discussing their upcoming book chapter, Lippstreu and Illingworth mentioned the potential benefits of the variety of information available through integrated talent acquisition and management systems. Everything from application zip code to key stroke characteristics to point-of-sale systems training results for new hires are housed together along with assessment results and a host of other details. This information can be extremely useful to practitioners, providing postimplementation consulting opportunities throughout the process (e.g. making adjustments to candidate sourcing practices, managing assessment content, modifying bands or cut scores, and demonstrating the assessment’s business impact).

Two related postimplementation applications of BD methods—predictor optimization and identification—are worth mentioning here as well. Some proponents see promise in applying data mining techniques to identify new ways to use existing predictors to enhance validity and to identify previously unknown or unexpected predictors. These methods have been put to good use in other professions and can also yield benefits in applied selection research. However, as previous articles have pointed out, a purely empirical, atheoretical approach to working with predictors is not considered a best practice (e.g., Maurath, 2014; Poeppelman et al., 2013). Cowen (2013) refers to data mining algorithms, or machine intelligences, that are capable of finding ways to capitalize on relationships in large, complex data sets that are incomprehensible to humans. It is important to ensure the modifications that come out of optimization efforts can be explained in a way that aligns with or builds on practitioners’ current understanding of predictor–criterion relationships (e.g. a positive linear, or upside-down u-shaped relationship between Conscientiousness and job performance; Carter, Dalal, Boyce, O’Connell, Chung & Delgado, 2014; Converse & Oswald,
A related key point underscored by Lippstreu and Illingworth is the importance of retaining links back to the job analysis. Data mining techniques can identify a host of unexpected, potentially useful patterns. But if the way the predictor is used alters dramatically, and the job has not changed, can the links to job analysis still be made?

Further, proponents—most likely those outside of our profession—might expect to find some kind of previously hidden game-changing performance predictor through BD and BDA. For this reason we strongly agree with authors of the previous related TIP articles that there are opportunities for I-Os to contribute to data science teams involved with talent acquisition. It is more likely that seemingly new predictors are actually indicators or proxies for one or more previously identified constructs, complete with their inherent concerns and limitations. The credit score is an illustrative example. It is an existing variable generated by an individual’s financial activity observed across vast data sources. The far-reaching effects of the housing crisis on personal debt have arguably compromised its reliability for the time being by introducing extreme real estate market volatility. However, although little validity research is available, preliminary evidence suggests the credit score can be a valid predictor of job performance. It is also probably a manifestation of previously known constructs such as conscientiousness and cognitive ability (Bernerth, Taylor, Walker, & Whitman, 2012). Finally, the sizable mean differences across demographic groups will certainly draw increased scrutiny for organizations choosing to use credit scores to screen candidates (Bernerth, 2012).

We suspect these conditions will apply to other seemingly new predictors identified via BD and BDA. It is unlikely that these techniques are going to resolve the diversity–validity dilemma (Pyburn, Ployhart, & Kravitz, 2008). If the “new” predictors they identify turn out to be composite manifestations of known characteristics, they will exhibit the same tendency for validity and group differences to vary together. In this case, end users are encouraged to make sure that group differences do not increase to levels beyond what would be expected from a more straightforward, well-understood predictor of comparable validity.

**Conclusion**

In sum, although a potential source of guidance for optimizing talent acquisition systems, BD volume is unnecessary for most applied selection research purposes at present. Likewise, we expect true BD velocity, particularly within-subjects velocity, to be of limited use on either the predictor or the criterion side. Finally, BD variety, like volume, can be a great source of insight regarding talent acquisition/management systems in general and could provide opportunities to incrementally improve or optimize selection programs in some settings. However, data mining will typically reveal proxies for constructs we have already encountered, and data-mining techniques must be used with care in predictor development in order for users to avoid increased risk. BD and BDA have captured our discipline’s interest and may end up being very useful in many ways. We acknowledge the potential for them to produce incremental contributions and opportunities for us to provide more insight as talent management...
consultants. Much of the enthusiasm in the context of applied selection research is justified, along with some caveats.

References


