A New Era of Large-Scale Data Sharing: A Test Publisher’s Perspective

JOHN W. JONES AND KELLY D. DAGES
Vangent, A General Dynamics Company

As part of a call to evaluate the trustworthiness of the industrial and organizational (I–O) psychology literature, Kepes and McDaniel (2013) make a request related to large-scale data sharing:

We encourage the release of data for meta-analyses in addition to data from primary studies. We note that past contentious debates concerning differing meta-analytic results in personality test validity (debates in the 1990s) and, more recently, in integrity test validity, would have been more clearly and reasonably resolved if the authors had released their data and related documentation.

As researchers within a test publisher organization, we have repeatedly been in the position where we have shared all available data for specific types of psychological tests (as was the case with Ones, Viswesvaran, & Schmidt, 1993), and we have also at times withheld sharing...
unpublished research\(^1\) (as was the case with Van Iddekinge, Roth, Raymark, & Odle-Dusseau, 2012). In relation to Ones et al., we shared published and unpublished research with significant and nonsignificant findings and replication studies. There are several factors that are considered when making a decision to share or not share. It is our belief that we are currently entering a new era of data sharing (see Table 1). This new era is impacted by five perspectives: (a) test publisher, (b) legal, (c) client organization, (d) academic, and (e) journal editor (Dages et al., 2013).

**New Era of Data Sharing**

It is our belief that many test publishers are increasingly embracing the new era of large-scale data sharing. In the past, most test publishers were entrepreneurial academics who went to market with a specific type of test. They owned their cottage test publishing businesses, which oftentimes provided supplemental income to their university careers. Today, many of the major test publishers have been purchased by large information management and technology firms that are publicly traded and are therefore regulated by the Securities Exchange Commission (SEC). These firms have large legal departments to protect the intellectual property that they purchased from test publishers, including large-scale datasets and proprietary research studies. These new owners are fully committed to making the investments necessary to continually improve their assessment products, technology platforms, and applied research processes. At the same time, however, they also require more rigor in terms of protecting all assets, including large-scale datasets. Given this new reality, large test publishers often find that responses to data requests from academic researchers are contingent upon the requirements and needs of legal departments and client organizations.

Legally, most data sharing requests to test publishers are now covered by binding contracts. In the age of “Big Data” and use of business analytics to gain competitive advantage (McAfee & Brynjolfsson, 2012), the value of data is at a premium. Vangent recently shared two large datasets with a university researcher that had a total estimated value of over $500,000. Contracts were needed to ensure that these data were protected, would not be shared with business competitors, and would be used only for the approved purpose. Had this contract not been signed, the data could not have been shared. These contractual requirements also raise concerns regarding Kepes and McDaniel’s recommendation of submitting raw data as part of the journal review process with the knowledge that the journal may publicly release that data at a future time.

Client organizations are also increasingly viewing their data as a source of strategic business advantage. Because of this, clients do not want their competitors to benefit from their local norms or validation studies. Nor do they want the public to know the quality of their talent applicant pool and overall passing rates. Therefore, contractual terms oftentimes exist between the client organization and test publishers as to who owns the data and whether or not datasets can even be shared.

Academics typically request all relevant studies and datasets for their research projects. The requesters range from advanced undergraduate students wanting to conduct a quality senior project to full professors who are considered experts in their fields. Many of these requesters want the data for large scale, definitive meta-analytic investigations, regardless of whether or not they are considered experts in this area as evidenced by a track record of quality publications. It is our experience that nearly all academics feel somewhat entitled to test publishers’ datasets, and

---

\(^1\) It is worth noting that test publishers do not control the sharing of published research nor do they withhold from sharing conference papers. The decision to share or not is typically centered on proprietary technical reports, confidential client validation studies, and proprietary datasets that have typically not been analyzed for publication purposes.
Table 1. Defining the New Era of Large-Scale Data Sharing

<table>
<thead>
<tr>
<th>Large-scale data sharing considerations</th>
<th>The Past Era</th>
<th>The New Era</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research requested</td>
<td>Published articles and conference papers</td>
<td>Published articles, conference papers, proprietary technical reports, and unpublished datasets</td>
</tr>
<tr>
<td>Dataset valuation</td>
<td>The overall valuation of the research and datasets was not seriously considered</td>
<td>In the age of business analytics and “Big Data” the valuation of the research and datasets is essential to protect these assets</td>
</tr>
<tr>
<td>Contractual requirements</td>
<td>Few, if any, contracts were used when sharing data with academic researchers</td>
<td>Contracts with academic researchers and/or their universities are becoming more commonplace in order to ensure quality research and to protect the shared assets</td>
</tr>
<tr>
<td>Client organizations</td>
<td>Clients typically used off-the-shelf assessment products and did not care if their data sets were shared by the publisher</td>
<td>Clients increasingly view their data sets as sources of strategic advantage and many do not want their data sets shared</td>
</tr>
</tbody>
</table>

A review of Table 1, which defines common large-scale data sharing considerations for the past era and how they are redefined for the new era of data sharing, suggests that in the past these academic researchers clearly had access to published studies and research presented at conferences. However, in the new era of data sharing the request has been broadened to include proprietary technical reports and even unpublished datasets. This type of request is very understandable in the digital age of “Big Data” and with the increased demand for predictive analytics (cf. McAfee & Brynjolfsson, 2012).

Journal editors clearly want to publish studies that are influenced scientifically by all relevant data. For instance, if journal editors are publishing a large-scale meta-analysis, they want to make sure that their journal guidelines for meta-analytic procedures are followed and that all current and available research is included. It is our perception that journal editors realize that their authors are immersed in a new era of data sharing and therefore these editors need to provide more editorials on well-functioning data sharing ecosystems to make sure that science never becomes restricted in any way.

Test Publisher Considerations

The transition from the old era to the new era of data sharing is surely still in progress, and some test publishers are on different points of this continuum than other publishers. However, test publishers who are members of trade associations like the Association of Test Publishers (ATP) are required to have PhD-level psychometricians on staff who strive to adhere to all relevant professional and legal standards. Moreover, these same publishers are fully committed to the highest quality of psychometric science when it comes to their assessment products. However, all publishers need to ask themselves a number of very important questions before they share their large-scale datasets with interested researchers. Many of these questions are designed to ensure that the research literature is not...
compromised due to inaccurate analyses (cf. Harris et al., 2012). These questions can help protect the trustworthiness of the research literature in addition to protecting the test publisher and client organizations. Still, as the new data sharing realities unfold, procedures need to be developed so that the criteria used to determine whether or not data can be released to a researcher is applied consistently and fairly across requesters, regardless of who is requesting the data. To that end, we have included a sampling of the types of decision-making questions that test publishers should routinely ask themselves before sharing large-scale datasets:

1. Does the researcher have scientific expertise in the area being investigated, and is there a track record of publications that confirms such expertise?
2. Is the research question well thought out, and can the researcher provide a credible proposal that the test publisher and their legal team can review?
3. If the researcher plans to conduct a meta-analytic investigation, is he/she highly competent with this type of research procedure?
4. Is the researcher approaching the study in a highly objective, scientific manner, or does this researcher have any known biases against the types of constructs that were measured by the types of tests used to generate the large-scale datasets requested?
5. Is the researcher bound by a set of professional and/or legal guidelines that govern his/her research efforts?
6. Are there any known conflicts of interest associated with sharing the dataset with the requesting researcher?
7. Are there any contractual obligations whereby the client organization(s) would need to approve of the sharing of the data with the interested researcher?
8. Is the researcher or his/her university open to signing a contract to protect the test publisher’s intellectual property and to prevent misuse of the dataset?
9. What quality control procedures does the researcher have in place to ensure that the datasets are accurately and securely coded, analyzed, and reported?
10. Finally, does the study require an unbudgeted investment by the publisher that the researcher could not reimburse (e.g., this could include dataset retrieval, compilation, formatting, checking, and packaging)?

References


