Conceptual and Methodological Profiling of Data: A Qualitative Metadata Process to Understand the Meaning and the Formation of Data Sources

Nathaniel J. Ratcliff, PhD and Joel Thurston, PhD

Social and Decision Analytics Division,
Biocomplexity Institute and Initiative, University of Virginia

Social and Decision Analytics Division
Biocomplexity Institute
University of Virginia

January 20, 2022

Funding: This research was sponsored by the US Army Research Institute for Behavioral and Social Science Research, cooperative agreement #W911NF2020027

Acknowledgments: We would like to thank the Social and Decision Analytics Division team who participated in the conceptual and methodological profiling and those who reviewed the manuscript.

Citation: Ratcliff N, Thurston J. Conceptual and Methodological Profiling of Data: A Qualitative Metadata Process to Understand the Meaning and the Formation of Data Sources, Technical Report. TR# 2023-284. Proceedings of the Biocomplexity Institute, University of Virginia; 2022 January. DOI: https://doi.org/10.18130/we5y-ez98.

Abstract

Data science research often involves ingesting and linking disparate sources of secondary data. While these sources can often be cleaned and wrangled into a usable form for analysis, robust documentation on how variables are created, and their intrinsic meaning, might not always be readily apparent. Without such meaning applied, data can lack the context necessary to understand the best ways to use and analyze it and risk misinterpretation. We introduce conceptual and methodological profiling processes into the data science pipeline as a qualitative tool to help researchers derive additional meaning and understanding from their data. Conceptual and methodological profiling uses various taxonomies to categorize variables and produce metadata to inform about how variables were created or recorded and the concepts they represent. To help explicate these processes, we first broadly describe these approaches and their place in the data science pipeline, then present a real-world example applying these techniques in our research using disparate data sources from the U.S. Army. Lastly, we discuss how researchers can find agreement while conducting these qualitative processes. We hope that the processes outlined here will provide data scientists additional tools to know their data better and how best to use it.

NOTE: The figures and tables can be found in the section titled "Figure Legends" towards the end of this document.

Introduction

"The numbers do not remember where they came from." —Lord (1953, p. 21)

Researchers often work with data from disparate sources, each with their unique provenance, formatting, and underlying structural layout. Though the pieces of data can often be cleaned and wrangled into a useable state for analysis, there is often no metadata that describes how they were created and what they are supposed to mean and represent. Without such meaning applied, the data may lack the context to determine the best way to use and analyze the data and risk misinterpretation. This paper introduces the process of conceptual and methodological profiling to provide researchers with tools to derive meaning and understanding from their data. This approach represents a tool, alongside other data management and fitness-for-use processes, for researchers using disparate primary and secondary/archival data sources to optimize their work. Below we outline a data science framework and introduce a conceptual and methodological data profiling process. Lastly, we present an in-depth walkthrough of conceptual and methodological profiling using real-world data sources from our ongoing projects with the U.S. Army.

A Framework for Doing Data Science

From the midst of the data revolution, data science has emerged as a transformative way to find meaning in our complex world (Provost & Fawcett, 2013; van der Aalst, 2020). Data science is an evolving field that transcends disparate methodological approaches (e.g., statistics, computer science), content areas (e.g., social, psychological, physical, geolocation), and levels of analysis (e.g., cellular, individuals, groups, nations; Garber, 2019; Wing, 2018). Data science often utilizes large, non-traditional forms of secondary data to draw insights into multifaceted problems (see Adjerid & Kelly, 2018; Keller et al., 2020; King et al., 2016). However, the data revolution is about more than just 'big data;' it's the joining of data of all sizes and types to address research questions that have never been answered before. As such, an organizing framework is needed to discover, access, repurpose, and statistically integrate all varieties of data—a *data science framework* (Keller et al., 2018; 2020; cf. discussion of the *data life cycle* in Berman et al., 2018). Using such a framework (see Figure 1), complex issues can be addressed to provide evidence-based insights via problem identification, data discovery, data ingestion & governance, and statistical analysis. Moreover, by creating standardized and repeatable processes, the data science framework guides the integration of disparate and novel data sources into research and ongoing analyses.

[insert figure 1 here]

Current Techniques for Profiling Data Quality and Usability

Data wrangling and assessment are a central part of the data science framework. Typically, once data have been ingested for a research project, researchers need to assess the quality and usefulness of the data for supporting analysis via data wrangling (see Keller et al., 2017; 2018; 2020). This process is iterative, first ensuring that all relevant data and associated metadata have been appropriately ingested. Next, to assess quality, the data are wrangled to

evaluate their timeliness, accuracy, geographic granularity, completeness, and reliability using various techniques (Dasu & Johnson, 2003; De Veaux et al., 2016; Wickham, 2014; Wing, 2019). This process is typically quantitative, focusing on errors, invalid values, outliers, and missing data points to help clean and transform the data for use in subsequent analyses. Our research found that these profiling techniques overlooked other essential aspects of the data that could be profiled to provide greater context and understanding of the data being used. Outlined below, we propose a new qualitative process that produces additional metadata integrated with traditional data profiling techniques, namely, *conceptual and methodological profiling*.

Assessing Qualitative Aspects of Data: Conceptual Profiling and Methodological Profiling

Raw numbers and text can only inform researchers about the state of being or *what is* about the data, not necessarily answering questions of meaning like the *for what* the data represent and *how* data were collected or recorded. In addition to profiling to assess the quality and format of data sources, it is essential to understand what concepts the variables represent and the methodology or process that produced the data (i.e., data provenance). Data provenance is a subject of increasing relevance to data science pipelines as a type of metadata that provides a contextual history of data and its relationship with data management systems (Doan et al., 2012; Glavic & Dittrich, 2007; Simmhan et al., 2005; Song et al., 2019). Over time, data can have a complex history involving numerous changes from its original source by being imported, transformed, or re-translated within and between data systems (Glavic & Dittrich, 2007). Along with other metadata, provenance provides the context to explain the origins of data which can build authenticity and trust in how to make sense of data and how it can be reused (Simmhan et al., 2005).

Importantly, contextual information is needed to guide how variables should be interpreted and used in subsequent analyses. Since data do not remember where they come from (Lord, 1953, p.21), data can be manipulated in any way that is mathematically feasible when conducting statistical analyses (e.g., addition, multiplication, regression) because these tests do not consider the objects or events to which the data refer. However, when it comes time for interpretation after an analysis, the question arises as to whether the results bear any meaningful relationship to the original objects or events being studied and thus, a conceptual/methodological issue arises rather than a statistical one (Howell, 2008, p. 21). Stated differently, results can be derived from a mathematically-sound statistical test, but this does not ensure that the methodology or conceptual meaning behind the test was sound or valid.

In the absence of pre-existing metadata and provenance, important conceptual and contextual information needs to be derived. Conceptual and methodological profiling provides a methodological framework for deriving this information and complements existing data profiling methods. These profiling processes can be performed in either order but are probably best done concurrently. Importantly, each categorization process is flexible and can be tailored to specific research needs by adding or subtracting the suggested qualitative taxonomies outlined below. Our social science research focuses on people, so our examples are related to data about individuals and groups. However, this process could be used for other data domains such as financial (e.g., stocks), non-human (e.g., animal behavior), physical (e.g., climate measurements), or mechanical processes (e.g., machine functioning).

Conceptual Profiling

Conceptual profiling involves deriving meaning from variables. It is a qualitative categorization process that involves identifying what constructs variables represent and measure conceptually. Constructs are latent, abstract (often theory-driven) conceptualizations representing ideas, experiences, and behaviors that can vary. As abstract concepts, constructs are unobserved (i.e., not directly measurable) and made concrete (indirectly) through the operationalization of observed measures or indicators (DeVellis, 2017; Kline, 2011). Researchers attempt to make constructs tangible and observable through an operationalization process where an observable or measurable variable is constructed to serve as an imperfect proxy for a construct (Morling, 2012; Pelham & Blanton, 2007; Stangor, 2015). For example, one might measure increases in heart rate and galvanic skin response (observable proxy) as a means of assessing a person's anxiety (latent construct).

In many cases, the observable or measurable variable is provided an operational definition that describes how it attempts to capture or define a construct in a concrete form. These observable measures can take many forms, from discrete categories (e.g., Did a person engage in a particular behavior: Yes or No) to nearly infinite magnitudes of scale (e.g., distance in meters a person traveled). Next, we describe some of the primary aspects of a construct that can be categorized to serve as metadata to inform future use and analysis (see Table 1). The aspects described below are not intended to be exhaustive or used in every research context. Different projects might require additional research-specific variable classifications depending on research aims.

Construct Identification. Determining what conceptual label can be applied to an observed variable in terms of what it represents, construct identification, is perhaps the most critical aspect of conceptual profiling. Observed variables are concrete representations of abstract constructs (DeVellis, 2017; Phelham & Blanton, 2007). Constructs may be identified based on the face validity of the variable name or description and the use of prior literature speaking to the intended purpose of the variable (e.g., research articles, surveys, forms). For example, a researcher might encounter a variable from a survey that includes a question about one's feelings of 'being down.' To describe yourself as feeling down is a colloquial means of expressing that you feel sad or depressed, so the construct attached to this variable might be 'depression.'

Researchers should use every available resource (e.g., the original survey or instrument used to collect the data, subject matter experts, contextual clues from other variables, other research that has used these data) to make a judgment as to what construct best captures what the variable represents. There may be cases in which multiple valid construct labels fit for a given variable. For example, you might attach both 'sadness' and 'depression' to the feeling down variable, at least initially. The use of multiple labels for a variable versus narrowing labels to a single construct will be determined by one's specific research needs. The use of multiple coders or judges can increase the agreement of the categorization process when multiple labels might fit (see Finding Agreement section below).

Construct identification also allows for researchers to apply a unification of terminology. In some cases, the raw data labels might have varying synonyms of terms between or within data sources that can be unified with a standard naming scheme during construct identification (e.g., consolidating variables labeled bereavement, grief, and sense of loss under a single term).

Another important factor to consider is the level of construct specificity. Construct terms can be applied to variables at various levels of specificity–from a particular instance to broad categories. Consider the following survey question: "Please indicate your current age in years." This item generally asks how old a person might be. A researcher could easily label the construct in this case as 'Person's Age,' which would capture the representation of the question. However, the researcher could use a more specific label for the construct such as 'Person's Age at Start of Term' or a more general label such as 'Demographics.' The relative sweet spot for the level of specificity is mainly dependent on the needs of the researcher and research questions. In some cases, it might be helpful to capture more than one level of construct specificity using multiple metadata labels (or tags) for the same variable. For example, one might be a higher-level construct label (e.g., Geospatial) and the other a more specific level (e.g., Home State).

Construct Span. The theoretical level of scaling of the construct. By definition, a variable must vary to some degree, so the span represents the degree to which a construct has any number of levels from two to infinity. Generally speaking, the span will fall into three major groupings: categorical, bounded rating scale, or continuous. *Categorical constructs* are those with nominal, discrete levels that can be dichotomous (e.g., Pass/Fail), unordered (e.g., a person's race), or ordered (e.g., rankings). *Bounded rating scales* are those with a limited number of levels that have standard rating scales (e.g., 5-point Likert scale of agreement). These scales can be unipolar (i.e., never to always) or bipolar (i.e., the contrast of two competing or opposing constructs at either end of a scale, like disagree to agree). Finally, fully *continuous scales* are those with an unlimited number of numeric levels (e.g., distance). Of note, the construct span might be at odds with how the construct is measured. For instance, a concept might be continuous but unnecessarily dichotomized (e.g., using a median split for age; see response types in Methodological Profiling).

Construct Referent. Describes *to whom* the construct is referencing or about at a given level of analysis (see Chan, 1998). Constructs and their reflective variable indicators vary in terms of the level of analysis in which they are operating, ranging from a singular entity to a vast system of interconnected entities and other, non-animate systems (e.g., weather). To determine at what level of analysis a construct or its reflective variable is operating, one can examine what referent is being used (Baltes et al., 2009; Field & Abelson, 1982; Glick, 1985; Klein et al., 2001). An *individual referent* refers to a singular entity (e.g., a person, a cell). For example, a survey question asking a response to "I often go to the park" uses an individual referent: the pronoun 'I.'. A *group referent* refers to a concept referencing more than one entity (e.g., work team, squad). For example, a survey question asking a response to "People on my team work hard" uses the group referent of 'people.' An *organizational referent* refers to a more extensive organizational system with many nested groups and individuals (e.g., a corporation, the Army). An *environmental referent* may refer to a concept that operates at a level beyond a single organization and affects many individuals, groups, and even organizations (e.g., policies, culture, weather, climate, geospatial landforms). Lastly, in some cases, a concept's referent might be ambiguous or mixed. An *ambiguous referent* is one where a referent cannot easily be determined (e.g., construct related to a timestamp). A *mixed referent* is one in which more than one level of analysis is being referenced (e.g., an aggregate group score using individual-level data).

Construct Form. Classifying the aspect of the referent entity that the construct is examining. The specific categories for this taxonomy may vary depending on research needs. However, the primary categories can be broken down into the following categories: characteristic, thought process, behavioral, biological, and index. A *characteristic form* is a concept that mainly describes an entity by a distinguishing feature (e.g., gender attribute, personality type, financial class). A *psychological form* is a concept that is related to a psychological or intrapersonal process happening in the mind of the entity that cannot be readily observed by others without the entity responding (e.g., self-report on a survey, cognitive ability test score). A *behavioral form* is related to an external, often interpersonal action taken by an entity that others can readily observe (e.g., number of times a person exercises a week, Yes/No responses to behavioral engagement or intent questions). Behavioral forms are typically related to actions that are intentionally under an entity's control. A *biological form* is a concept that pertains to a process or aspect of an entity's biological systems (e.g., heart rate, cholesterol level). It is usually not directly under conscious control. An *index form* type refers to a combination of the disparate component forms or concepts within forms mentioned above (e.g., health vulnerability might be a combination of a person's education, food access, and health diagnoses; see Bollen & Bauldry, 2011).[1] Not every construct form will be applicable to every type of referent. For example, organizations do not typically have psychological or biological processes associated with them, but indexes could be constructed aggregating psychological or biological information across multiple individual members of an organization.

Construct Framework. Classifying constructs regarding how they might be used in a conceptual or statistical modeling framework. The first two categories concern categorizing potential predictors as either trait-like or state-like (cf. Steyer et al., 2015). *Trait-like predictors* are characteristics of an entity that are relatively stable or take long periods of time to change (e.g., a person's race, personality type). *State-like predictors* are flexible constructs and time-varying (e.g., heart rate, person's age). *Situational predictors* are distal to an entity and provide a contextual backdrop that can be controlled (e.g., state of residence, university course section number). *Outcomes* are constructs representing a result or end-state that can determine the relationship with some cause or predictor construct (e.g., attrition, performance evaluation).

Operational Intent. The description provided by the creator of the measured variable. This description provides context as to the original intent for defining and using the variable. If none is provided by data documentation or prior research, the researcher may provide a definition based on other available conceptual information (e.g., secondary publications referencing the data). Note that that the operational intent provided by the creator of a measured variable may not necessarily be valid due for a number of reasons (e.g., failed measurement validation, imprecise wording of items). This is why it is important for researchers to gather as much information as possible about the data they are using (see Additional Resources section).

Concept Importance. The concept's relative importance or usefulness to a researcher's needs. This concept can be accomplished using a simple rating system to evaluate constructs' use towards a given research aim (e.g., 5-point rating scale ranging from 1 = *not important* to 5 = *very important*).

---

[1] This type contrasts with composite measures, which are single variables representing a combination (either additively or through an average) of interchangeable items that are all reflective of the same underlying latent construct.

Additional Resources. Other resources that may help researchers better understand the meaning behind constructs and measured variables include (a) obtaining additional literature that describes the data (e.g., empirical articles, published and unpublished reports); (b) descriptions of data collection attributes (e.g., population, coverage, repeated measurement); (c) original documentation (e.g., forms, surveys, data collection protocols); and (d) measure development and validation results. The more information that can be collected about the data under consideration the better, because it is often examining the totality of the evidence that provides the most clarity in identifying the construct(s) associated with a given variable.

[insert Table 1 here]

Methodological Profiling

Measurement is a fundamental quality of science that can be defined as the assignment of values to an object in such a way as to correspond to different degrees of a quality or property of some object, person, or event (Duncan, 1984; Stevens, 1946). Thus, for measurement to occur, three things are necessary (Albano, 2017): (a) one needs an *object or thing* that is being measured (in matters of social and public policy, this is often people); (b) a variable for which a *property or quality* is being measured for an object (i.e., a construct); and (c) a *value or units* in which measurement is captured within a variable (i.e., concrete assessment). How measured variables represent abstract concepts can take on many approaches using numerous measurement instruments (for a review, see DeVellis, 2017). Importantly, the decisions made in the development of measured variables have downstream consequences to the inferences drawn from them in subsequent analyses. Moreover, not all measurement applications are created equal; the concrete way in which an abstract construct is represented can be slightly imprecise at best due to an inherent degree of measurement error involved in measurement. At worst, the measurement of abstract concepts can be invalid or misleading. As the statistician George Box put it, "All models are wrong, but some are useful" (Box & Draper, 1987, p. 424). The usefulness of a measure in a statistical analysis is largely determined by the degree it represents what it was intended to measure (i.e., construct validity).

Methodological profiling involves deriving how variables were formed. It is a qualitative categorization process that involves identifying how variables were created by understanding the process used in their measurement and recording. Methodological profiling often involves some form of data sleuthing to uncover the origins of variables. Much like conceptual profiling, researchers should examine the codebooks, original forms (e.g., record forms/documents), survey instruments, and research articles that speak to how variables were measured, formulated, and recorded to glean this information.

Next, we describe some of the primary aspects of measurement that can be categorized to serve as metadata to inform future use and analysis (see Table 2). Again, as discussed with conceptual profiling, the aspects described below are not intended to be exhaustive or applicable to every research context—usage and terminology will vary by project.

Data Type. How data were created or obtained, covering the fundamental data types that underlie most data applications (see Keller et al., 2017; 2018; 2020). *Administrative data* are collected for primarily administrative use within an organization, program, or service process

(e.g., health records, property tax data). *Designed data* have traditionally been used in scientific discovery as they result from an intentional process to observe and collect data (e.g., surveys, experiments, remote sensing). *Opportunity data* are derived from Internet-based information collected unobtrusively through websites or apps via application programming interfaces (APIs) and web-scraping methods. *Procedural data* focus on the documented processes and policies within organizations and governments (e.g., laws, standard operating procedures).

Observation Source. How closely the measured data are to whom is being observed. A *direct source* is where the entity being measured primarily provides the data through their actions or a direct result of their actions (e.g., self-report survey, interview text, blood pressure reading). An *indirect source* is where the entity being measured has another secondary or intermediary entity providing the data about the entity of focus (e.g., administrative data, leader assessment, health care provider diagnosis). In some cases, this determination might not be readily apparent, and thus, an 'unknown' categorization is appropriate.

Measure Occasion. The degree to which a measure is collected once or repeatedly over different time occasions. If the measure is repeated, the frequency of repeated measurements should be captured in terms of the total number (e.g., once, twice, thrice) or the frequency of the measurements (e.g., daily, weekly, monthly, yearly, varying).

Item Stem. Text that provides the context for orienting a response to a question or statement. The context may be temporal (e.g., "In the last four weeks…"), situational (e.g., "When going out with friends…"), locational or geospatial (e.g., "At home…"), etc. Item stems are common with designed data and are often found on surveys and forms as the same prefatory clause paired with multiple, different items (e.g., "In the last four weeks, how often have you consumed alcohol? In the last four weeks, how often have you smoked marijuana? In the last four weeks, how often have you used narcotics?").

Item Text. The exact, word-for-word text used to describe a question or statement that is often found on surveys and forms (e.g., "How often do you exercise?", "I often clean my room."). For some types of data (e.g., administrative, procedural) the item text may simply be a short phrase qualifying a measurement (e.g., "Blood Pressure," "Race," "Birth Date"). Having the exact textual wording provides an unfiltered look at how a variable was measured.

Item Number. Refers to the positioning or ordering of a variable being asked in a more extensive set (e.g., item #24 on a survey or form). This information can help determine the possibility of ordering effects.

Response Format. Describes how a variable was measured or recorded. The response format taxonomy includes the following categories: (a) *dichotomous* where only two discrete response options are provided (e.g., Yes | No); (b) *categorical* where three or more discrete response options are provided (e.g., Education Level: High School | College | Graduate); (c) *bounded rating scale* where a rating scale is used (e.g., 5-point Likert scale from 1 = *Never* to 5 = *Always*); (d) *composite* where multiple, interchangeable items reflecting the same construct are aggregated into a single value (e.g., via summation or averaging) to form a composite variable (e.g., depression scale based on averaging 23 items); (e) *index* where multiple, unrelated items

reflecting disparate constructs are combined (via weighting, averaging) to form a new construct in a single value of a new variable (e.g., socioeconomic status formed from the components of income, education, and job type); (f) *free response* where any number of numerical or textual values can be responses (e.g., open-ended questions, numerical age in years); and, (g) *date* where responses are a date in some combination of indices of time (e.g., 2010-10-24, 12 January 2004).

Response Values. The actual values that were available for the measured variable. For *dichotomous and categorical variables*, values consist of a list of categories that correspond to a code (e.g., 'M,' '0 = Male'). For *bounded rating scales*, values correspond to a point on a rating scale that may or may not have labeled anchors (e.g., 1 = *Unimportant*, 2 = *Neutral*, 3 = *Important*). For *composite scales, indexes, and numeric free responses*, values represent a number on a continuous scale at varying levels of precision (e.g., 23, 100.1, 44.56). For *textual free responses*, values are simply a character-for-character record of what was written, dictated, or typed. Lastly, for *dates*, a range of event dates or timestamps are recorded in a specific format (e.g., POSIXct format: '2012-01-23').

Expected Range. Determining what is reasonably expected for numeric data in terms of minimum and maximum values. For instance, if a variable measures a person's age, one would expect the values to range from above zero to around 122 (age of the oldest person on record). Values that fall outside this range could be identified for further scrutiny and classified as invalid if no other explanation can be provided.

Measure Quality. The quality or trustworthiness of the measured variable. Information related to the methodology of measurement, or the formation of the data, can be used to evaluate its relative quality (e.g., reports on measure development and validation, copies of surveys and forms). Item text can provide insight into possible quality issues based on how questions were phrased on surveys and forms. Particularly with designed and administrative data, issues such as grammatical errors, double-barreled questions, response options with restricted range, social desirability, high sensitivity, order effects, survey fatigue, or practice effects can all introduce error or noise into the data collection process. Measure quality can be assessed using simple categorization (e.g., Low, Average, High) or using a rating scale assessing quality (e.g., 5-point Likert scale: 1 = *Low Quality* to 5 = *High Quality*). Assessing quality can be a particularly subjective experience. Reviewing as much additional information about the variables as possible (e.g., order in relation to other variables from the same source), leveraging subject matter expertise in survey design and experimental methods, and ensuring agreement across multiple raters will help ensure accurate quality assessments.

Measure Source. The originating source of a measured variable. Here, it is essential to document the original source of the variable by name (e.g., an item from the Values in Action-Inventory of Strength scale) and provide a relevant citation (e.g., Peterson, 2007). Documenting the original sources (or, if possible, acquiring a copy such as a pdf of a survey) will provide researchers with primary source information about a variable and the methodology used to generate it.

Additional Resources. Other resources that may help researchers better understand the methodology behind measured variables include: (a) additional literature that describes the

formation or use of the measured variable (e.g., empirical articles, published, unpublished reports); (b) measure development and validation results; (c) original documentation (e.g., forms, surveys, data collection protocols); and, (d) annotated comments on any issues observed regarding the measured variable (e.g., odd values/codes, grammatical errors, duplicate variables).

[insert Table 2 here]

Data Table Profiling

Data table profiling uses a mix of conceptual and methodological profiling techniques to describe the entire data table in which the individual variables are housed. This profiling step allows categorizing data at the aggregate, table-level to quickly understand the data contained within using similar conceptual and methodological profiling taxonomies as listed above (see Table 3). Some of this information may be the same or similar to other common types of metadata collected and disseminated with data sets.

Predominant Construct. The predominant construct being captured within the data table at a high level of abstraction. For example, a data table reflecting items on a health questionnaire could be labeled 'Health Records' to describe the entirety of the data.

Predominant Data Type. The predominant data type using the data type categories described above (e.g., administrative, designed, opportunity, and procedural). For example, a table reflecting data collected from a survey on unit climate could be categorized as predominately 'designed' in nature.

Predominant Referent. The table's predominant data type using the construct referent categories described above (e.g., individual, group, organizational, environment, mixed/ambiguous).

Data Owner. Who is or was the data owner when it was collected or received? For example, the data owner for the American Community Survey (ACS) is the U.S. Census Bureau. However, in more ambiguous cases, like scraped data from a website or app, the website or app could be listed as the owner (i.e., source) or unknown.

Represented Population. Captures whom the data represent in terms of a population targeted or sampled. For example, a survey conducted on Soldiers entering the Army could have 'Active Duty Army Soldiers' as the represented population for the data table.

Data Time Frame. The period covered by the data in the table from its earliest point in time to its latest. This designation typically requires the data table to contain variables with some sort of filing or event dates or be associated with some other external information speaking to the time frame covered by the data (e.g., earliest date: '2000-11-22'; latest date: '2019-07-15').

Data Update Frequency. How often the data table is updated (e.g., daily, weekly, monthly, quarterly, annually, or decennially). In some instances, as with experimental datasets,

the data might represent a one-time collection with a designation of 'one-time' or 'never' being an appropriate categorization of the data's frequency.

Linkable Data. Indicates whether individual or group identifiers are contained within the data table that could be used to link to other data tables (e.g., full names, social security numbers, researcher-created identifier codes).

Identifiable Information. Describes whether there is individually identifiable data contained within the data table. This information might include variables that directly identify individuals (e.g., full names, social security numbers, identification numbers). Alternatively, one might also indicate identifiable information is present if the data set is comprehensive enough that multiple variables could be combined to identify specific individuals with some confidence (e.g., age + gender + zip code + race + birth month).

Data Sensitivity. The degree to which the data table contains sensitive data and thus, should have various levels of restricted access. *Low sensitivity* data have little sensitive information (e.g., publicly available data tables) and can be made openly accessible to any interested researcher. *Moderate sensitivity* data have some information that should not be freely distributed to the public and might require limiting access to those who meet specific criteria upon request of the data (e.g., data collected from a survey). Lastly, *highly sensitiv*e data can be damaging to individuals if misused (e.g., data containing personal identifiers, classified data). These data should have restricted access only to named individuals who have proper oversight approvals (e.g., Institutional Review Board, data owner).

Ethical Procurement. Describes whether the data are ethically obtained or sourced. For example, using survey data obtained via informed consent would meet the ethical procurement criteria. By contrast, data scraped from a user's social media account without their knowledge could be considered unethical. In other cases where the data source is not transparent, a determination might be challenging and labeled 'unknown' or not used.

Additional Resources. Other resources that may provide researchers with important information or context about a data table include providing an annotated description for the data table describing its purpose, history, and any issues with its usage.

[insert Table 3 here]

Finding Agreement

Aspects of conceptual and methodological profiling are qualitative and, thus, somewhat subjective. Therefore, it can be important to ensure a certain level of agreement for the application of typologies when profiling variables. The best way to demonstrate agreement is to have multiple (two or more) independent raters (sometimes called judges) profile the variables using a standard set of typologies and categories. Depending on the size of the data source, raters can either make judgments for the entire corpus of data or a select (preferably random) subset of variables (e.g., 10% of the total variables). It is important that raters all use the same number of categories or scales. To develop a standard set of typologies and categories, rater may review and

discuss an initial sampling from the data set to establish common reference points. However, raters should make their judgments of the remainder of the corpus or their assigned subset independent from one another. Once all raters have profiled the data, there are statistical tests to determine whether a sufficient level of agreement has been reached between raters (for more details, see Determining Levels of Agreement below). Instances with high disagreement can be resolved with further discussion.

List of Best Practices

We have outlined a list of best practices researchers can use during the conceptual and methodological profiling process as they seek to better understand the data they are using. Again, not all points will apply to every research project.

Conceptual Profiling

- Obtain theoretical review papers discussing the concepts being measured and defined as well as the operational intent of the measures.
- Identify a primary construct for every variable; note that some variables may be associated with multiple constructs.
- Use a level of specificity for typological categories that works best for research needs.
- Document the units of measurement for a given variable (e.g., categorical, continuous) and identify the referent for the measured variables (e.g., an individual, group, environment).

Methodological Profiling

- When possible
    - obtain provenance about the data source along with all relevant metadata;
    - obtain documentation of measure validation (e.g., results, reports, published articles);
    - obtain original forms, surveys, and online scripts to provide context for how data were collected (e.g., formatting, item wording, ordering, response options);
    - obtain complete codebooks describing unique categories along with their codes as well as suggested weighting schemes;
    - obtain intended scaling and composite variable formation for rating scales, including whether certain variables should be reverse-scored.
- Synthesize information for variables and scales that have gone through multiple iterations documenting significant changes over time (i.e., version changes).
- Make notes of any irregularities or errors that might affect the interpretation of variables (e.g., marking double-barreled questions).

Data Table Profiling

- Understand the population that was targeted or was likely a passive data provider in the universe of data collection or scraping.
- Document the period of time over which the data was collected and frequency with which it was collected.
- Indicate whether identifiable information is present and whether the data can be linked to other data sources.

Metadata Documentation and Agreement

- Compile all metadata generated from conceptual and methodological profiling in a knowledge portal (e.g., database, spreadsheet) linked to the variables for easy searching and filtering by researchers.
- Ask additional researchers conceptually and methodologically profile the data sources (either all or a subset) to determine levels of agreement.

Data Profiling in Action: A Real-World Example

Research Project Overview

The conceptual and methodological profiling of data sources described below was performed as a part of a collaborative research project between the Biocomplexity Institute of the University of Virginia and the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI).[2] This research effort explores how the U.S. Army can derive insights from large amounts of disparate data sources. The project goal is to examine the feasibility of using data analytics to predict performance by Soldier characteristics in the U.S. Army (see Figure 2) using the large amount of available administrative data collected to support mission effectiveness (National Academies of Sciences, Engineering, and Medicine, 2017, 2019).

[insert Figure 2 here]

Materials and Methods

Data Sources Profiled

We requested and gained access to 23 data tables, including personnel records, training, entry screenings, and physical fitness tests (see Table 4). The source is the Army's Person-Event Data Environment (PDE). The PDE is a secure, remote-access, virtual data enclave that provides access to Army data, including psychological measures, performance indicators, medical information, and administrative personnel records across the careers of individual Soldiers (National Academies of Sciences, Engineering, and Medicine, 2017; Vie et al., 2015). The Army Analytics Group (AAG) supports the PDE, and the Research Facilitation Laboratory (RFL) administers the PDE. Use of the PDE is available to researchers and institutions that wish to conduct research using Army and Department of Defense (DOD) data sources provided proper approvals are met (Knapp et al., 2018). Before starting the research, researchers obtain approvals for the study from the University of Virginia Institutional Review Board and the Army Human Research Protections Office. In addition, researchers must apply for and receive a military or DOD Common Access Card to access and use the data.

[insert Table 4 here]

[2] Cooperative agreement #W911NF-19-2-0164: "The Social Component of the Human Dimension: Leveraging Existing DOD Data Towards Optimized Individual and Team Performance in the Army."

The PDE provides a limited set of metadata describing variables similar to a variable codebook (see Table 5 for examples). These variable descriptions include the data table name in which the variable resides (ENT_NAME), the name of the variable (VAR_NAME), a descriptive name of the variable (VAR_BUSNAME), a description of what the variable represents or the question being asked (VAR_DESCRIPTION), and an entry providing further commentary on how to use the variable or, in some cases, a listing of response options categories (VAR_USAGE). Crosswalks of categorical codes were also available from different repositories within the PDE (e.g., 'PDE_LOOKUP').

When coding schemes were not available, we sought out this information from the data owners. We identified additional contextual information for variables by searching repositories containing the original forms and surveys used to collect the data, as well as published articles documenting the use of data sources (e.g., validation efforts of measures). Once the available metadata and contextual information was gathered, we began the conceptual and methodological profiling.

[insert Table 5 here]

Conceptual and Methodological Profiling

The lead author conceptually and methodologically profiled 3,179 variables across the 23 data tables provisioned in the PDE.[3] The lead author recorded the PDE metadata and the different typologies to be conceptually and methodologically profiled within a single spreadsheet, then profiled each variable by assigning a response for each of the 18 conceptual and methodological categories listed below. The spreadsheet served as a central metadata repository with each variable represented as a row and the profiling categories as columns (see Figure 3).

[insert Figure 3 here]

Construct Identification. A single word or phrase (e.g., Depression, Physical Activity, Date, Person's Race) describing the construct identified based on PDE-derived metadata and any available external information (e.g., original data collection source document).

Construct Referent. The referent level of analysis of the construct; a categorical variable with five values:

- *Individual* (a single individual);
- *Unit* (a group of individuals);
- *Environment* (a place or larger societal context);
- *Mixed* (a mix of two or more categories mentioned above or ambiguous);
- *NA* (typology not applicable to construct).

Construct Form. The aspect of the entity the construct examines; a categorical variable with eight values:

- *Attribute* (relatively stable characteristic, e.g., Person’s Race);

[3] As a preliminary check of consistency, the second author profiled a random 10% subset of the same data, which yielded similar results upon comparison. For more rigorous testing of agreement, see the agreement section.

- *Personality* (stable personality type, e.g., Extraversion);
- *Cognitive* (mental or cognitive ability, e.g., SAT Score);
- *Perceptual* (self-assessed perception of the self, others, or the environment, e.g., Mood);
- *Behavioral* (description of actual or intended behavior, e.g., Exercise Frequency);
- *Biological* (a physical or biological indicator, e.g., Blood Pressure);
- *Index* (a multi-dimensional construct, e.g., SES);
- *NA* (typology not applicable to construct).

Construct Framework. How the construct might be used in our conceptual framework, given our project's focus on identifying indicators of performance; a categorical variable with five values:

- *Situational* (an external factor that influences the entity, e.g., post location);
- *Trait* (an internal characteristic that stays relatively constant across time, e.g., personality, cognitive ability);
- *State* (an internal characteristic that tends to change over time, e.g., emotions, age);
- *Outcome* (externally observable end state of interest, e.g., attrition, causality); and
- *Performance* (defined as a work behavior or action that Soldiers engage in to further the goals of the organization, e.g., work quality/quantity, helping co-workers).

Performance Type. For variables identified as performance-related, we further categorized them using four performance dimensions identified by Koopmans and colleagues (2011), as well as a fifth general performance category we created:

- *Task Performance* (related to proficiency on central job tasks, e.g., work quantity and quality);
- *Contextual Performance* (behaviors that support organizational goals outside direct tasks like showing initiative or helping co-workers; cf. organizational citizenship behavior; Organ, 1967);
- *Counterproductive Performance* (actions that harm the well-being of the organization, e.g., absenteeism, substance abuse);
- *Adaptive Performance* (the degree to which individuals adapt to changing work roles, e.g., problem-solving, learning new tasks);
- *General Performance* (for variables that reflected an overall indicator of performance across multiple dimensions);
- *NA* (typology not applicable to construct).

Data Table Name. A short name given to a data table to describe it (e.g., Entry Table for data table containing records upon entry to the Army).

Data Source. The data owner or organization that produced the data table (e.g., the data source for the Entry Table is the Military Entrance Processing Command or MEPCOM).

Data Type. The type or form of data (cf. Keller et al., 2018); a categorical variable with four values:

- *Administrative* (data derived from the operation of administrative systems for record-keeping, transaction, or registration, e.g., demographics, health tests);
- *Designed* (data that have traditionally been used for scientific discovery and meant for research purposes to capture some sort of concept, e.g., self-esteem, cognitive tests);
- *Opportunity* (data which are generated on an on-going basis as society moves through its daily paces, e.g., geolocation, social media, fitness sensors);
- *Procedural* (data derived from laws, procedures, regulations or manuals, e.g., Uniformed Code of Military Justice).

Item Stem. For items (i.e., questions or statements) from surveys or forms, the stem is the prefatory text that provides specific context for the body of the item the respondent is referencing with when making a judgement or providing a response (e.g., 'In the last two weeks…').

Item Text. For items (i.e., questions or statements) from surveys or forms, the item's text that requires the respondent to make a judgment or to provide a response (e.g., '…I have felt happy').

Item Number. Where the item appeared on a survey or form; its order relative to the other items from the same source (e.g., Question 34).

Operational Intent. The operational definition of the variable provided by the creator of the variable or defined in prior research. If this information was unavailable, a researcher provided a definition based on available construct information.

Response Format. How the variable was measured or recorded using:

- *Dichotomous* (having two discrete categories, e.g., polar yes-no questions);
- *Categorical* (having more than two discrete categories, e.g., Person Race);
- *Bounded Rating Scale* (a rating scale with multiple options of increasing or decreasing intensity, e.g., degree of disagreement);
- *Composite* (a composite average or sum of other items, e.g., a depression scale averaging five questions about depressive symptoms);
- *Free Response* (an unbounded entry of text or numeric data, e.g., indicate age in years);
- *Date* (the date of an event).

Response Values. A listing or labeling of the responses including:

- Nominal values with corresponding categorical code labels (e.g., 1 = Yes, 0 = No or White | Black | Asian | Other);
- Scale points with corresponding labeled anchors (e.g., -1 = Disagree, 0 = Neutral, 1 = Agree);
- *Text* for entry of words or sentences (e.g., for home city: ‘Columbus’);
- *Numeric* if the values are a continuous set of numbers (e.g., current age in years: 19);
- *Event date* for dates in various formats (e.g., ‘2010-09-15’, ‘10-24-2009’).

Reverse Coded. Indicates whether an item should be reversed-coded when creating composite scale variables (i.e., 'Yes' or 'No').

Source Scale. The name of the scale an item originates from if previously developed for a composite scale (e.g., Values in Action-Inventory of Strengths).

Citation. The name of the report or publication in which the variable was first described (e.g., Peterson, 2007).

Profiler's Comments. A catch-all for any issues or use cases for the profiled variable. For example, if a survey question asks about two topics as one question (double-barreled questions) or was dropped from a later version of a survey, it would be noted here.

An example of outputs from the conceptual and methodological profiling process for the variables presented in Table 5 can be found in Tables 6 and 7, respectively. For a detailed description of the example typology classifications and the justifications thereof, please see the Supplemental Information.

[insert Table 6 here]

For conceptual profiling, we identified the conceptual characteristics for each variable and the operational intent of the variable.[4] For example, the variable representing a survey item about shoplifting (ID#5) was identified as 'Theft' at the individual level, reflects a behavior the individual performs, could be used in a construct to measure counterproductive behavior within our performance framework, and is defined to measure incidences of stealing (see Table 6).

[insert Table 7 here]

For methodological profiling, we identified the methodological characteristics for each variable regarding how variables were measured or reported. For example, the variable representing a survey item about shoplifting (ID#5) was identified as a *Designed* variable, using a *Dichotomous* 'Yes' or 'No' (nominal) measurement scale (see Table 7).

Determining Levels of Agreement

A set of five independent judges categorized the same random subset of 156 variables drawn from all data sources, representing about 5% of the original corpus profiled by the single researcher.[5] For purposes of determining agreement, the judges used the following six typologies: Construct Identification, Construct Referent, Construct Form, Construct Framework, Performance Type, and Data Type. All but the construct identification typologies involved choosing from among a limited set of categories (i.e., for construct reference there were five

[4] When available, the operational intent was taken from the original data collectors in the form of expressed definitions or extracted from variable descriptions. In the absence of any information attached to a variable, a general operational intent was assessed for the variable by the secondary data researchers.

[5] Though we used five judges as an in-lab exercise, two to three judges should be sufficient with three easily breaking ties. We felt that, given the size of the corpus of variables profiled, a 5% subset was sufficient for validation.

options: Individual, Unit, Environment, Mixed, and NA). However, the free-response nature of construct identification holds the potential for raters to generate many different synonyms (e.g., anger, rage, aggression) and levels of specificity (e.g., anger vs. negative affect), which exponentially lowers the likelihood of establishing any kind of agreement across judges. Therefore, for construct identification, the five judges picked from a pre-generated list of 70 possible constructs. This list represented all constructs identified by the original judge when profiling the 156 variables. Judges participated in a one-hour training to become familiar with the different typology classifications, practice profiling selected items, and allow for discussion to establish a shared understanding of the task. The total estimated time for completing the task was 2–3 hours per judge (for an example of items judged, see the Supplemental Information).

Interrater agreement (IRA) is typically assessed using Fleiss's kappa (Fleiss, 1971; Fleiss et al., 2003) in cases for which there are more than two judges.[6] Kappa (κ) is generally a better measure than a simple percent agreement, as κ takes into account the possibility of the agreement occurring by chance. Fleiss' kappa proposes three critical assumptions: (a) judgments should be categorical (either nominal or ordinal); (b) judges should use the same categories, and (c) the judges are independent of one another. The following guidelines are given for interpreting values of Fleiss's Kappa (see Fleiss et al., 2003): .00–.40 = poor agreement beyond chance; .40–.75 = fair to good agreement beyond chance; > .75 = excellent agreement beyond chance.

Results

Fleiss's Kappa was calculated for each of the six typologies categorized by the five judges, as well as an overall average indicator of agreement (see Table 8). Results indicated fair to excellent agreement for the different typologies (kappas = .41 to .70) with all significantly exceeding chance levels. Construct identification and construct reference typologies provided the relatively strongest levels of agreement with kappas of .70 and .69, respectively (i.e., good agreement). Choosing constructs from a list of terms and identifying the subject of the construct seemed to be the easiest for judges. Construct framework and performance type had the relatively weakest levels of agreement with kappas of .41 and .45 (i.e., fair agreement). Classifying a predictor-type variable as a changing state or stable trait seemed to be the most challenging determination. Overall, the agreement was within acceptable ranges and provided validity for the subjective classifications.

[insert Table 8 here]

Conclusion

Taken together, conceptual and methodological profiling helps researchers identify the meaning behind the variables they are working with and how best to use them in further modeling and analysis. We hope that the profiling processes described here will provide researchers additional tools to know their data better and how best to use it. A deeper understanding of data yields better modeling usage, ultimately providing more sound and nuanced inferences to the research queries being assessed.

---

[6] Cohen's kappa can be used for cases where only two judges make qualitative judgments (see Cohen, 1960). In contrast, quantitative judgments by raters (magnitude of a rating) should be made using interval scales. They would require different metrics of IRA such as intraclass correlation (ICC), r_{WG}, and a_{WG} (for a review, see LeBrenton & Senter, 2008).

Acknowledgments

The authors thank Sallie Keller, Stephanie Shipp, Aaron Schroeder, and Joanna Schroeder for their comments on a draft of this manuscript. Also, the authors thank Alyssa Mikytuck, Josh Goldstein, Eric Oh, and Kathryn Linehan for their assistance with interrater agreement portion of this research. Research in this article has not been published or disseminated previously.

Author Funding and Disclosure Statements

This research was supported by a cooperative agreement (# W911NF1920164) between the University of Virginia and the U.S. Army Research Institute for the Behavioral and Social Sciences. The authors declare that there are no potential conflicts of interest with respect to the research, authorship, funding, and/or publication of this article.

References

Adjerid, I., & Kelley, K. (2918). Big data in psychology: A framework for research advancement. *American Psychologist*, *73*, 899–917. https://doi.org/10.1037/amp0000190

Albano, T. (2017). *Introduction to educational and psychological measurement: Using R*. Available from https://cehs01.unl.edu/aalbano/intro-measurement-r/

Baltes, B. B., Zhdanova, L. S., & Parker, C. P. (2009). Psychological climate: A comparison of organizational and individual level referents. *Human Relations*, *62*, 669–700. https://doi.org/10.1177/0018726709103454

Berman, F., Rutenbar, R., Christensen, H., Davidson, S., Estrin, D., Franklin, M., Hailpern, B., Martonosi, M., Raghavan, P., Stodden, V., & Szalay, A. (2016). Realizing the potential of data science: Final report from the National Science Foundation Computer and Information Science and Engineering Advisory Committee Data Science Working Group. Retrieved from https://www.nsf.gov/cise/ac-data-science-report/CISEACDataScienceReport1.19.17.pdf

Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological Methods*, *16*, 265–284. https://doi.org/10.1037/a0024448

Box, G. E. P., & Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*. Wiley.

Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). *A Theory of Performance*. In Schmitt, N. Borman, W., and Associates (Eds.). Personnel Selection in Organizations (pp. 35–70). Jossey−Bass Publishers, San Francisco.

Campbell, J. P., & Wiernik, B. M. (2015). The modeling and assessment of work performance. *The Annual Review of Organizational Psychology and Organizational Behavior*. *2*, 47–74. https://doi.org/10.1146/annurev−orgpsych−032414−111427

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology*, *83*, 234–246. https://doi.org/10.1037/0021-9010.83.2.234

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46. https://doi.org/10.1177/001316446002000104

Dasu, T., & Johnson, T. (2003). *Exploratory data mining and data cleaning*. Wiley.

De Veaux, R., Hoerl, R., & Snee, R. (2016). Big data and the missing links. *Statistical Analysis and Data Mining*, *9*, 411–416. https://doi.org/10.1002/sam.11303

DeVellis, R. F. (2017). *Scale Development: Theory and Applications* (4th ed., Vol. 26). Sage Publications.

Doan, A., Halevy, A., & Ives, Z. (2012). Data provenance. *Principles of data integration* (pp. 359–371). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-416044-6.00014-4

Duncan, O. D. (1984). *Notes on social measurement: Historical and critical*. Russell Sage.

Field, R. H. G., & Abelson, M. A. (1982). Climate: A reconceptualization and proposed model. *Human Relations*, *35*, 181–202. https://doi.org/10.1177/001872678203500302

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*, 378–382. https://doi.org/10.1037/h0031619

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical Methods for Rates and Proportions* (3rd ed.). Wiley-Interscience.

Garber, A. M. (2019). Data Science: What the Educated Citizen Needs to Know. *Harvard Data Science Review, 1.1,* 1–14. https://doi.org/10.1162/99608f92.88ba42cb

Glavic, B., & Dittrich, K. R. (2007). Data provenance: A Categorization of existing approaches. In: Kemper, A., Schöning, H., Rose, T., Jarke, M., Seidl, T., Quix, C., & Brochhaus, C. (Eds.), *Datenbanksysteme in Business, Technologie und Web (BTW) / 12. Fachtagung des GI-Fachbereichs Datenbanken und Informationssysteme (DBIS)* (pp. 227–241). Aachen. Bonn.

Glick, W. H. (1985). Conceptualizing and measuring organizational and psychological climate: Pitfalls in multilevel research. *Academy of Management Review*, *10*, 601–616. https://doi.org/10.2307/258140

Howell, D. C. (2008). *Fundamental Statistics for the Behavioral Sciences* (6th ed.). Thomson Wadsworth.

Keller, S., Korkmaz, G., Orr, M., Schroeder, A., & Shipp, S. (2017). The evolution of data quality: Understanding the transdisciplinary origins of data quality concepts and approaches. *Annual Review of Statistics and Its Application*, *4*, 85–108. https://doi.org/10.1146/annurev-statistics-060116-054114

Keller, S., Korkmaz, G., Robbins, C., & Shipp, S. (2018). Opportunities to observe and measure intangible inputs of innovation: Definitions, operationalization, and examples. *PNAS*, *115*, 12638–12645. https://doi.org/10.1073/pnas.1800467115

Keller, S. A., Shipp, S. S., Schroeder, A. D., & Korkmaz, G. (2020). Doing data science: A framework and case study. *Harvard Data Science Review*, *2.1*, 2–28. https://doi.org/10.1162/99608f92.2d83f7f5

Klein, K. J., Conn, A. B., Smith, D. B., & Sorra, J. S. (2001). Is everyone in agreement? An exploration of within-group agreement in employee perceptions of the work environment. *Journal of Applied Psychology*, *86*, 3–16. https://doi.org/10.1037/0021-9010.86.1.3

Kline, R. B. (2011). *Principles and Practice of Structural Equation Modeling* (5th ed., pp. 3–427). The Guilford Press.

King, E. B., Tonidandel, S., Cortina, J. M., & Fink, A. A. (2016). Building understanding of the data science revolution and I-O psychology. In: Tonidandel, S., King, E. B., & Cortina, J. M. (Eds.), *Big Data at Work: The Data Science Revolution and Organizational Psychology* (pp. 1–15). Routledge. https://doi.org/10.4324/9781315780504

Knapp, D., Asch, B. J., DeMartini, C., Ruder, T., & Hanley, J. M. (2018). *Using the Person-Event Data Environment for military personnel research in the department of defense: An evaluation of capability and potential uses*. RAND Corporation: Santa Monica, CA. https://www.rand.org/pubs/research_reports/RR2302.html

Koopmans, L., Bernaards, C. M., Hildebrandt, V. H., Schaufeli, W. B., de Vet, H. C. W., & van der Beek, A. J. (2011). Conceptual frameworks of individual work performance: A systematic review. *Journal of Occupational and Environmental Medicine*, *53*, 856–866. https://doi.org/10.1097/JOM.0b013e318226a763

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*, 815–852. https://doi.org/10.1177/1094428106296642

Lord, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist*, *8*, 750–751. https://doi.org/10.1037/h0063675

Morling, B. (2012). *Research Methods in Psychology: Evaluating a World of Information*. W.W. Norton & Company.

National Academies of Sciences, Engineering, and Medicine. (2017). *Strengthening Data Science Methods for Department of Defense Personnel and Readiness Missions*. National Academies Press.

Organ, D. W. (1988). *Organizational Citizenship Behavior: The Good Soldier Syndrome*. Lexington Books.

Peterson, C. (2007). *Brief Strengths Test*. Cincinnati: VIS Institute.

Phelham, B. W., & Blanton, H. (2007). *Conducting Research in Psychology: Measuring the Weight of Smoke* (3rd ed.). Thomson Wadsworth.

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data*, *1*, 51–59. https://doi.org/10.1089/big.2013.1508

Simmhan, Y. L., Plale, B., & Gannon, D. (2005). A survey of data provenance in e-science. *ACM SIGMOD Record*, *34*, 31–36. https://doi.org/10.1145/1084805.1084812

Song, J., Alter, G., & Jagadish, H. V. (2019). C2Metadata: Automating the capture of data transformations from statistical scripts in data documentation. *In Proceedings of the 2019 International Conference on Management of Data (SIGMOD '19)*. *Association for Computing Machinery*. https://doi.org/10.1145/3299869.3320241

Stangor, C. (2015). *Research Methods for the Behavioral Sciences* (5th ed.). Cengage Learning.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677–680. https://doi.org/10.1126/science.103.2684.677

Steyer, R., Mayer, A., Geiser, C., & Cole, D. A. (2015). A theory of states and traits--revised. *Annual Review of Clinical Psychology*, *11*, 71–98. https://doi.org/10.1146/annurev-clinpsy-032813-153719

van der Aalst, W. M. P. (2020). The data science revolution: How learning machines changed the way we work and do business. In: Strous, L., Johnson, R., Grier, D. A., & Swade, D. (Eds.), *Unimagined futures: ICT opportunities and challenges* (pp. 5–19). Springer. https://doi.org/10.1007/978-3-030-64246-4_2

Vie, L. L., Scheier, L. M., Lester, P. B., Ho, T. E., Labrthe, D. R., Seligman, M. E. P. (2015). The U.S. Army Person-Event Data Environment: A military-civilian big data enterprise. *Big Data*, *3*, 1–13. https://doi.org/10.1089/big.2014.0055

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*, 1–23. https://doi.org/10.18637/jss.v059.i10

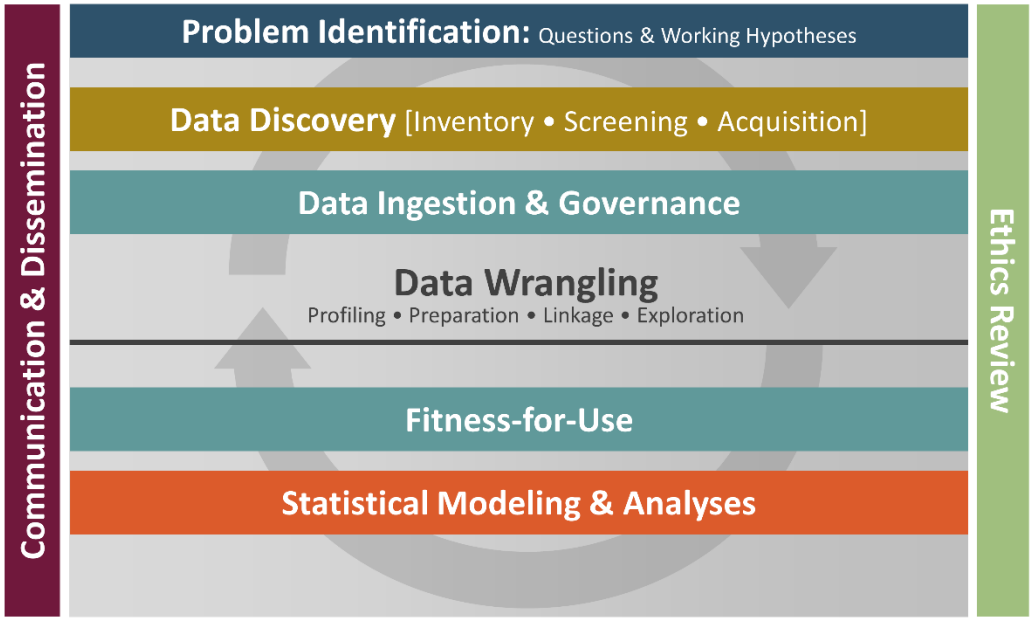
Wing, J. M. (2019). The data life cycle. *Harvard Data Science Review, 1*, 1–6. https://doi.org/10.1162/99608f92.e26845b4

Wing, J. M., Janeja, V. P., Kloefkorn, T., & Erickson, L. C. (2018). Data Science Leadership Summit: Summary Report. Technical Report. National Science Foundation, USA. https://doi.org/10.13140/RG.2.2.13710.61764

Figure Legends

Figure 1

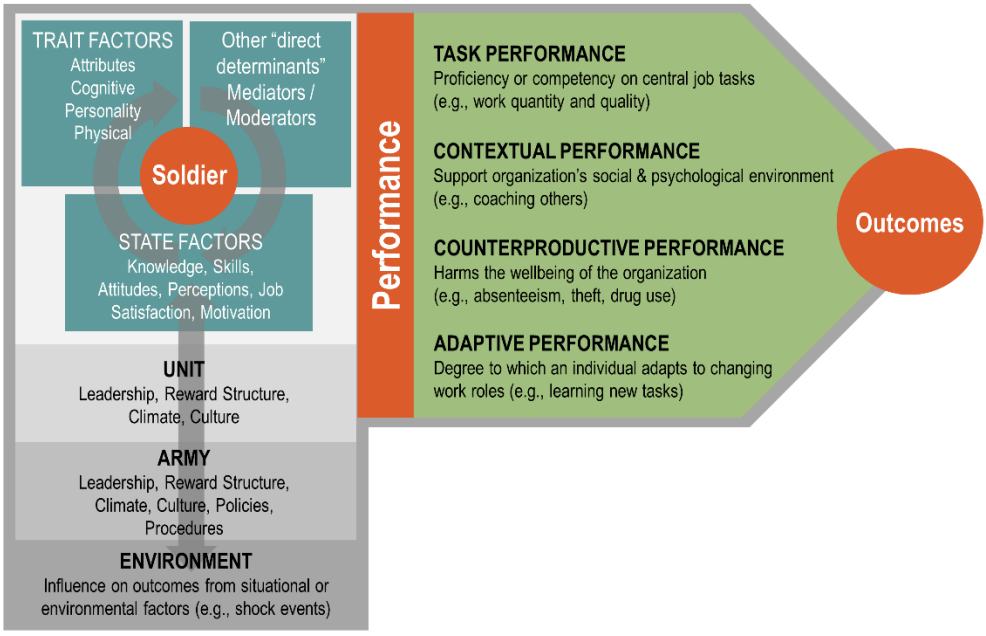
Data Science Framework



Note. The data science framework starts with the research question, or problem identification, and continues through the following steps: data discovery—inventory, screening, and acquisition; data ingestion and governance; data wrangling—data profiling, data preparation and linkage, and data exploration; fitness-for-use assessment; statistical modeling and analyses; communication and dissemination of results; and ethics review (Keller et al., 2020).

Figure 2

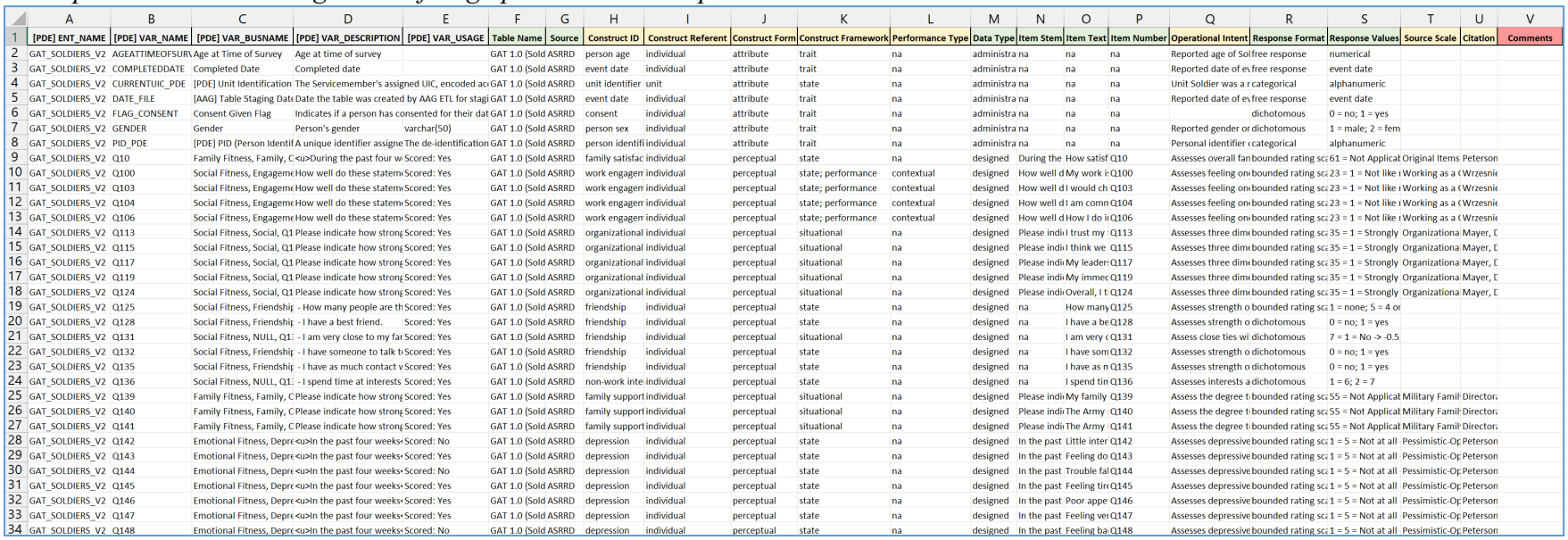
Hierarchal Conceptual Performance Framework



Note. The conceptual performance framework is derived from a synthesis of Army and academic literature on individual and teamwork performance (cf. Campbell et al., 1993; Campbell & Wiernik, 2015; Koopmans et al., 2011).

Figure 3

Conceptual and Methodological Profiling Spreadsheet Example



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	[PDE] ENT_NAME	[PDE] VAR_NAME	[PDE] VAR_BUSNAME	[PDE] VAR_DESCRIPTION	[PDE] VAR_USAGE	Table Name	Source	Construct ID	Construct Referent	Construct Form	Construct Framework	Performance Type	Data Type	Item Stem	Item Text	Item Number	Operational Intent	Response Format	Response Values	Source Scale	Citation	Comments
2	GAT_SOLDIERS_V2	AGEATTIMEOFSUR	Age at Time of Survey	Age at time of survey		GAT 1.0 (Sold	ASRRD	person age	individual	attribute	trait	na	administra	na	na	na	Reported age of Sol	free response	numerical			
3	GAT_SOLDIERS_V2	COMPLETEDDATE	Completed Date	Completed date		GAT 1.0 (Sold	ASRRD	event date	individual	attribute	trait	na	administra	na	na	na	Reported date of ev	free response	event date			
4	GAT_SOLDIERS_V2	CURRENTUIC_PDE	[PDE] Unit Identification	The Servicemember's assigned UIC, encoded ac		GAT 1.0 (Sold	ASRRD	unit identifier	unit	attribute	state	na	administra	na	na	na	Unit Soldier was a r	categorical	alphanumeric			
5	GAT_SOLDIERS_V2	DATE_FILE	[AAG] Table Staging Date	Date the table was created by AAG ETL for stagi		GAT 1.0 (Sold	ASRRD	event date	individual	attribute	trait	na	administra	na	na	na	Reported date of ev	free response	event date			
6	GAT_SOLDIERS_V2	FLAG_CONSENT	Consent Given Flag	Indicates if a person has consented for their dat		GAT 1.0 (Sold	ASRRD	consent	individual	attribute	trait	na	administra	na	na	na		dichotomous	0 = no; 1 = yes			
7	GAT_SOLDIERS_V2	GENDER	Gender	Person's gender	varchar(50)	GAT 1.0 (Sold	ASRRD	person sex	individual	attribute	trait	na	administra	na	na	na	Reported gender or	dichotomous	1 = male; 2 = fem			
8	GAT_SOLDIERS_V2	PID_PDE	[PDE] PID (Person Identif	A unique identifier assigne	The de-identification	GAT 1.0 (Sold	ASRRD	person identifi	individual	attribute	trait	na	administra	na	na	na	Personal identifier (	categorical	alphanumeric			
9	GAT_SOLDIERS_V2	Q10	Family Fitness, Family, C	<u>During the past four w	Scored: Yes	GAT 1.0 (Sold	ASRRD	family satisfac	individual	perceptual	state	na	designed	During the	How satisf	Q10	Assesses overall fan	bounded rating sc	1 = Not Applicat	Original Items	Peterson	
10	GAT_SOLDIERS_V2	Q100	Social Fitness, Engageme	How well do these statem	Scored: Yes	GAT 1.0 (Sold	ASRRD	work engagem	individual	perceptual	state; performance	contextual	designed	How well d	My work is	Q100	Assesses feeling on	bounded rating sc	3 = 1 = Not like	Working as a (	Wrzesnie	
11	GAT_SOLDIERS_V2	Q103	Social Fitness, Engageme	How well do these statem	Scored: Yes	GAT 1.0 (Sold	ASRRD	work engagem	individual	perceptual	state; performance	contextual	designed	How well d	I would ch	Q103	Assesses feeling on	bounded rating sc	3 = 1 = Not like	Working as a (	Wrzesnie	
12	GAT_SOLDIERS_V2	Q104	Social Fitness, Engageme	How well do these statem	Scored: Yes	GAT 1.0 (Sold	ASRRD	work engagem	individual	perceptual	state; performance	contextual	designed	How well d	I am comn	Q104	Assesses feeling on	bounded rating sc	3 = 1 = Not like	Working as a (	Wrzesnie	
13	GAT_SOLDIERS_V2	Q106	Social Fitness, Engageme	How well do these statem	Scored: Yes	GAT 1.0 (Sold	ASRRD	work engagem	individual	perceptual	state; performance	contextual	designed	How well d	How I do it	Q106	Assesses feeling on	bounded rating sc	3 = 1 = Not like	Working as a (	Wrzesnie	
14	GAT_SOLDIERS_V2	Q113	Social Fitness, Social, Q1	Please indicate how strong	Scored: Yes	GAT 1.0 (Sold	ASRRD	organizational	individual	perceptual	situational	na	designed	Please indi	I trust my	Q113	Assesses three dim	bounded rating sc	5 = 1 = Strongly	Organizationa	Mayer, D	
15	GAT_SOLDIERS_V2	Q115	Social Fitness, Social, Q1	Please indicate how strong	Scored: Yes	GAT 1.0 (Sold	ASRRD	organizational	individual	perceptual	situational	na	designed	Please indi	I think we	Q115	Assesses three dim	bounded rating sc	5 = 1 = Strongly	Organizationa	Mayer, D	
16	GAT_SOLDIERS_V2	Q117	Social Fitness, Social, Q1	Please indicate how strong	Scored: Yes	GAT 1.0 (Sold	ASRRD	organizational	individual	perceptual	situational	na	designed	Please indi	My leader	Q117	Assesses three dim	bounded rating sc	5 = 1 = Strongly	Organizationa	Mayer, D	
17	GAT_SOLDIERS_V2	Q119	Social Fitness, Social, Q1	Please indicate how strong	Scored: Yes	GAT 1.0 (Sold	ASRRD	organizational	individual	perceptual	situational	na	designed	Please indi	My immed	Q119	Assesses three dim	bounded rating sc	5 = 1 = Strongly	Organizationa	Mayer, D	
18	GAT_SOLDIERS_V2	Q124	Social Fitness, Social, Q1	Please indicate how strong	Scored: Yes	GAT 1.0 (Sold	ASRRD	organizational	individual	perceptual	situational	na	designed	Please indi	Overall, I t	Q124	Assesses three dim	bounded rating sc	5 = 1 = Strongly	Organizationa	Mayer, D	
19	GAT_SOLDIERS_V2	Q125	Social Fitness, Friendshi	- How many people are th	Scored: Yes	GAT 1.0 (Sold	ASRRD	friendship	individual	perceptual	state	na	designed	na	How many	Q125	Assesses strength o	bounded rating sc	1 = none; 5 = 4 or			
20	GAT_SOLDIERS_V2	Q128	Social Fitness, Friendshi	- I have a best friend.	Scored: Yes	GAT 1.0 (Sold	ASRRD	friendship	individual	perceptual	state	na	designed	na	I have a be	Q128	Assesses strength o	dichotomous	0 = no; 1 = yes			
21	GAT_SOLDIERS_V2	Q131	Social Fitness, NULL, Q1	- I am very close to my far	Scored: Yes	GAT 1.0 (Sold	ASRRD	friendship	individual	perceptual	situational	na	designed	na	I am very c	Q131	Assess close ties wi	dichotomous	7 = 1 = No -> -0.5			
22	GAT_SOLDIERS_V2	Q132	Social Fitness, Friendshi	- I have someone to talk t	Scored: Yes	GAT 1.0 (Sold	ASRRD	friendship	individual	perceptual	state	na	designed	na	I have som	Q132	Assesses strength o	dichotomous	0 = no; 1 = yes			
23	GAT_SOLDIERS_V2	Q135	Social Fitness, Friendshi	- I have as much contact v	Scored: Yes	GAT 1.0 (Sold	ASRRD	friendship	individual	perceptual	state	na	designed	na	I have as n	Q135	Assesses strength o	dichotomous	0 = no; 1 = yes			
24	GAT_SOLDIERS_V2	Q136	Social Fitness, NULL, Q1	- I spend time at interests	Scored: Yes	GAT 1.0 (Sold	ASRRD	non-work inte	individual	perceptual	state	na	designed	na	I spend tin	Q136	Assesses interests a	dichotomous	1 = 6; 2 = 7			
25	GAT_SOLDIERS_V2	Q139	Family Fitness, Family, C	Please indicate how strong	Scored: Yes	GAT 1.0 (Sold	ASRRD	family support	individual	perceptual	situational	na	designed	Please indi	My family	Q139	Assess the degree t	bounded rating sc	55 = Not Applicat	Military Famil	Director	
26	GAT_SOLDIERS_V2	Q140	Family Fitness, Family, C	Please indicate how strong	Scored: Yes	GAT 1.0 (Sold	ASRRD	family support	individual	perceptual	situational	na	designed	Please indi	The Army	Q140	Assess the degree t	bounded rating sc	55 = Not Applicat	Military Famil	Director	
27	GAT_SOLDIERS_V2	Q141	Family Fitness, Family, C	Please indicate how strong	Scored: Yes	GAT 1.0 (Sold	ASRRD	family support	individual	perceptual	situational	na	designed	Please indi	The Army	Q141	Assess the degree t	bounded rating sc	55 = Not Applicat	Military Famil	Director	
28	GAT_SOLDIERS_V2	Q142	Emotional Fitness, Depre	<u>In the past four weeks<	Scored: No	GAT 1.0 (Sold	ASRRD	depression	individual	perceptual	state	na	designed	In the past	Little inter	Q142	Assesses depressive	bounded rating sc	1 = 5 = Not at all	Pessimistic-Op	Peterson	
29	GAT_SOLDIERS_V2	Q143	Emotional Fitness, Depre	<u>In the past four weeks<	Scored: Yes	GAT 1.0 (Sold	ASRRD	depression	individual	perceptual	state	na	designed	In the past	Feeling do	Q143	Assesses depressive	bounded rating sc	1 = 5 = Not at all	Pessimistic-Op	Peterson	
30	GAT_SOLDIERS_V2	Q144	Emotional Fitness, Depre	<u>In the past four weeks<	Scored: No	GAT 1.0 (Sold	ASRRD	depression	individual	perceptual	state	na	designed	In the past	Trouble fal	Q144	Assesses depressive	bounded rating sc	1 = 5 = Not at all	Pessimistic-Op	Peterson	
31	GAT_SOLDIERS_V2	Q145	Emotional Fitness, Depre	<u>In the past four weeks<	Scored: Yes	GAT 1.0 (Sold	ASRRD	depression	individual	perceptual	state	na	designed	In the past	Feeling tir	Q145	Assesses depressive	bounded rating sc	1 = 5 = Not at all	Pessimistic-Op	Peterson	
32	GAT_SOLDIERS_V2	Q146	Emotional Fitness, Depre	<u>In the past four weeks<	Scored: Yes	GAT 1.0 (Sold	ASRRD	depression	individual	perceptual	state	na	designed	In the past	Poor appe	Q146	Assesses depressive	bounded rating sc	1 = 5 = Not at all	Pessimistic-Op	Peterson	
33	GAT_SOLDIERS_V2	Q147	Emotional Fitness, Depre	<u>In the past four weeks<	Scored: Yes	GAT 1.0 (Sold	ASRRD	depression	individual	perceptual	state	na	designed	In the past	Feeling wo	Q147	Assesses depressive	bounded rating sc	1 = 5 = Not at all	Pessimistic-Op	Peterson	
34	GAT_SOLDIERS_V2	Q148	Emotional Fitness, Depre	<u>In the past four weeks<	Scored: No	GAT 1.0 (Sold	ASRRD	depression	individual	perceptual	state	na	designed	In the past	Feeling ba	Q148	Assesses depressive	bounded rating sc	1 = 5 = Not at all	Pessimistic-Op	Peterson	

Note. The example spreadsheet captures the integration of metadata from the conceptual and methodological profiling process from the real-world example. PDE = Person-Event Data Environment, ID = identification.

Tables

Table 1

Table of Conceptual Profiling Typologies

Construct Typology	Description	Question Answered	Example Categories or Metadata
Construct Identification	Identifies the latent construct that is best represented by an observable variable.	What conceptual label can be applied to an observed variable?	- Gender Code → Person's Gender - Question asking about group dynamics → Group Climate - Question about feeling down → Depression
Construct Span	Describes the number of levels of a construct.	How is the construct typically scaled?	- Dichotomous, Categorical, Bounded Rating Scale, Continuous
Construct Referent	Describes at what level of analysis the construct operates.	Whom is the construct about?	- Individual, Group, Organization, Environment, Mixed
Construct Form	Describes what aspect of entity or thing the construct examines.	What aspect of an entity does the construct examine?	- Characteristic, Psychological, Behavior, Biological, Index - Attribute, Cognitive, Perceptual, Personality, Behavior, Biological, Index - Characteristics, Psychological, Affective, Social, Educational, Economic
Construct Framework	Describes how the construct might be used in a conceptual or statistical modeling framework.	How might the construct fit into a conceptual framework or be modeled statistically?	- Trait, State, Situational, Outcome
Operational Intent	Describes the operational definition of the variable provided by the creator or prior research.	How do the original creators define the variable?	- Quoted definitions or descriptions of constructs and measured variables
Construct Importance	Describes the relative importance of the construct to the research question or purpose of the data collection.	How important is this construct or variable?	- 1 = not at all important; 2 = somewhat important; 3 = moderately important; 4 = important; 5 = very important
Additional Resources	Identifies external sources of information related to the construct and measured variable.	What external information is available to contextualize the concepts measured in the data?	- Empirical articles, reports, writings, validations

Table 2

Table of Methodological Profiling Typologies

Measure Typology	Description	Question Answered	Example Categories or Metadata
Data Type	Identifies the type of data as it relates to its original collection intent.	How was the data created or obtained?	- Administrative, Designed, Opportunity, Procedural
Observation Source	Identifies the source of the observation of data obtained or recorded.	How closely is the measured data to the entity being observed?	- Directly, Indirectly, Unknown - Primary, Secondary, Unknown
Measure Occasion	Identifies whether a measure is a one-time event or measured over repeated time occasions.	How often is the data measured?	- One-time, Repeated, Continuously - Once, Twice, Three Times, Four Times - One-Time, Repeated Daily, Repeated Weekly, Repeated Annually, Repeated Varying
Item Stem	Text identifying the context of the statement or question.	What is the context of the item response?	- "In the last four weeks…" - "While in class…"
Item Text	Text identifying the central question or statement seeking a response.	What is the central text of the question or statement?	- "How often do you exercise?" - "I can pay attention without distractions."
Item Number	Identifies the number or positioning of the statement or question in a larger set.	Where did the variable fall in an ordering of variables?	- 34, Q24, Question 334, Item 5
Response Format	Identifies how the variable was measured or recorded.	What format are the values of the variable?	- Dichotomous, Categorical, Bounded Rating Scale, Composite, Index, Free Response, Date
Response Values	Identifies the values that were available or are represented for the measured variable.	What are the values of the variable?	- List of all category codes and related labels, the listing of scale points and labeled anchors, numeric, event date
Expected Range	Identifies the expected range of values for a measured numeric variable.	What are the expected minimum and maximum values that are reasonable for a numeric variable?	- Age in months: 1 to 1,464
Measure Quality	Identifies the degree of quality of the measured variable.	What is the quality of the variable?	- Low, Average, High - Rating Scale: 1 = Very Low Quality, 2, 3, 4, 5 = Very High Quality
Measure Source	Identifies the source where the variable originated.	Where did the variable come from?	- Peterson, C. (2007). *Brief Strengths Test*. Cincinnati: VIS Institute.
Additional Resources	Identifies external sources of information related to the measured variables and any variable issues.	What external information is available to contextualize how the variables were measured?	- Obtaining original forms, surveys, validation reports - Comments on issues with the variables

Table 3

Table of Data Table Profiling Typologies

Measure Typology	Description	Question Answered	Example Categories or Metadata
Predominant Construct	Identifies the predominant construct being captured within the data table at a high level of abstraction.	What conceptual category is predominately represented within the data table?	- Health Records, Fitness Records, Demographics
Predominant Data Type	Identities the predominant data type represented within the data table.	What data type predominantly exists within the data table?	- Administrative, Designed, Opportunity, Procedural
Predominant Referent	Identifies the predominant level of analysis represented by variables within the data table.	What is the predominant referent of the variables within the data table?	- Individual, Group, Organization, Environment
Data Owner	Identifies the original creator or provider from which the data table was sourced.	Who is the owner of the data table?	- U.S. Census Bureau, Defense Manpower Data Center
Representative Population	Identifies the target population that the data in the table is sourced from.	What population was the data sampled from?	- College students, U.S. Population older than 16 years old, Active Duty Army Soldiers
Data Time Frame	Identifies the time frame of coverage for the data table.	What time period do the data cover?	- 2000-04-12 to 2018-11-24; a period of two weeks
Data Update Frequency	Identifies the frequency with which the data in the data table are updated.	How often is the data table updated?	- Never (One-Time), Daily, Weekly, Monthly, Annually
Identifiable Information	Identifies whether the data table contains personal identifiers or enough information for the identification of individuals or groups.	Does the data table contain identifiable information?	- Yes, No
Linkable Data	Identifies whether the data table can be linked to other data tables using codes or identifiers.	Can the data table be linked to other data tables?	- Yes, No
Data Sensitivity	Identifies the level of sensitivity of the data within the data table.	What is the degree of sensitivity of the data?	- Low, Moderate, High; Open, Limited, Restricted
Ethical Procurement	Identifies the degree to which the data were ethically sourced.	Was the data ethically obtained?	- Ethical, Unethical, Unknown
Additional Resources	Identifies external sources of information related to the data table and annotates issues of usage with the data.	What external information is available to contextualize how the data table was created?	- Obtaining original forms, surveys, validation reports - Comments on overall issues with usage of the data table

Table 4

Table of Person-Event Data Environment (PDE) Data Sources Profiled

Table Name	Data Source Name	PDE Table Name	Description	# Vars	PID Availability	First Date	Last Date
Master	Active Duty Military Personnel Master	MV_MASTER_AD_ARMY_QTR_V3A	Master administrative records	161	Yes	2001-09-30	2019-12-31
MEPCOM	Military Entrance Processing Command	MEPCOM_USAREC_RA_ANALYST	Initial entry records	124	Yes	2000-10-01	2016-07-19
OMAHA 5	Supplemental Health Questionnaire OMAHA 5	MEPCOM_OMAHA5_201605	Entry behavioral health screening questionnaire	60	Yes	2000-01-01	2019-07-09
Transaction	Active Duty Military Personnel Transaction	MV_TRANS_AD_ARMY_30_V3A	Entry and exit status within the Army	44	Yes	2001-09-01	2018-12-31
TAPAS	Tailor Adaptive Personnel Assessment	DMDC_TAPAS_201602	Personality test for placement upon entry	155	Yes	2010-03-01	2015-05-01
Training 1	Individual Training History	MV_INV_TRN_HIST_ARMY	Records of courses and training classes taken	12	Yes	1978-04-01	2018-04-01
Training 2	Army Training and Requirements Resource System	TA_ATRRS	Records of course information and completion status	29	Yes	1978-11-15	2018-5-22
Training 3	Digital Training Management System	TA_DTMS_TRAINING	Records of training classes taken and completed	9	Yes	2001-01-01	2016-06-30
Weapon Qual	Digital Training Management System	TA_DTMS_WEAPON_QUAL	Records of weapons qualification training	12	Yes	2001-01-01	2016-06-23
APFT	Army Physical Fitness Test	TA_DTMS_APFT	Records of physical fitness test scores	22	Yes	2001-01-21	2016-06-13
Height/Weight	Height & Weight	TA_DTMS_HT_WT	Records of height and weight test	13	Yes	2001-01-15	2016-06-13
GAT 1.0	Global Assessment Tool (Active Duty Soldier)	GAT_SOLDIERS_V2	Survey assessment of psychosocial characteristics	132	Yes	2009-05-05	2014-01-29
GAT 2.0	Global Assessment Tool (Active Duty Soldier)	GAT_SOLDIERS_20_V2	Survey assessment of psychosocial characteristics	210	Yes	2013-09-09	2017-09-30
URI	Unit Risk Inventory	ARDSURV_URI2_201602	Survey screening for high-risk behaviors and attitudes in units	68	No	2002-03-01	2016-12-05
URI-R	Reintegration Unit Risk Inventory	ARDSURV_URIR3_201603	Survey screening for high-risk behaviors and attitudes in units during deployment or post-deployment	103	No	2008-10-16	2016-03-10
DEOCS	Defense Organizational Climate Survey	DEOMI_DEOCS_ARMY_MIL	Survey on unit issues related to effectiveness, equal opportunity, and sexual assault response & prevention.	160	No	2014-04-13	2016-09-30
PHA 1	Medical Operational Data System	TA_PHA_OLDFORM_V1	Records of periodic health assessment	355	Yes	1982-09-10	2017-04-01
PHA 2	Medical Operational Data System	TA_PHA_NEWFORM_V25	Records of periodic health assessment	611	Yes	2007-12-01	2017-03-20
Pre-DHA 1	Medical Operational Data System	TA_DHA_DD2795_199905	Records of pre-deployment health assessment	57	Yes	2002-11-13	2013-03-15
Pre-DHA 2	Medical Operational Data System	TA_DHA_DD2795_201209	Records of pre-deployment health assessment	162	Yes	2012-10-22	2017-04-03
Post-DHA 1	Medical Operational Data System	TA_DHA_DD2796_200801	Records of post-deployment health assessment	392	Yes	2003-01-22	2013-03-15
Post-DHA 2	Medical Operational Data System	RWJF_DHA_DD2796_200801	Records of post-deployment health assessment	255	Yes	2008-01-18	2013-03-15
Derogatory Statements	Interactive Personnel Elective Records Management System	TA_IPERMS_DEROG_V2	Records of negative papers and statements	9	Yes	2001-01-01	2018-06-16
Awards	Army Work Force Transaction File	MV_AWTF_AWARDS	Records of awards received	24	Yes	2012-03-28	2018-12-31

Note. PID = Person Identifier; # Vars = Number of variables. Total number of tables profiled = 23; total number of variables profiled = 3,179.

Table 5

Examples of Unmodified Metadata for Variables in the Person-Event Data Environment (PDE)

Example ID#	ENT_NAME	VAR_NAME	VAR_BUSNAME	VAR_DESCRIPTION	VAR_USAGE
1	MV_MASTER_AD_ARMY_QTR_V3A	ADSVC_PE_DT	Active Duty Service Projected End Date	The date for which a DoD Military Service member is projected to leave Active Service. For Enlisted only, also referred to as Enlisted Active Service Projected End Date or ETS of Minimum Service. For Officers only, also referred to as Expected Active Duty End Date	Before October 2000, this date applied to enlisted only, and the officer date was stored in Enlisted Active Service Obligation End or Officer Active Status Projected End Date. Applicable only to enlisted members.
2	MEPCOM_USAREC_RA_ANALYST	RECORD	Record Status	Current record status	NA
3	GAT_SOLDIERS_V2	Q47	Emotional Fitness, Character, Q47	Think about how you have acted in actual situations <u>during the past four weeks</u>. Please answer only in terms of what YOU actually did. Please read carefully. Select a number from 0 to 10 according to how often you showed/used the qualities listed? - Prudence or caution	Scored: Yes
4	ARDSURV_URI2_201602	UIC_PDE	[PDE] Unit Identification Code	The Servicemember's assigned UIC is encoded according to PDE data security procedures.	NA
5	ARDSURV_URI2_201602	Q36	Criminal History, Q36	Within the past 12 months, have you stolen or shoplifted anything	Scored: Yes

Note. NA = no data provided.

Table 6

Conceptual Profiling of Example Metadata

Example ID#	Construct Identification	Construct Referent	Construct Form	Construct Framework	Performance Type	Operational Intent
1	Date	Individual	Attribute	Trait	NA	The date for which a DoD Military Service member is projected to leave Active Service. For Enlisted only, also referred to as Enlisted Active Service Projected End Date or ETS of Minimum Service.
2	Admin	NA	NA	NA	NA	Reported administrative information.
3	Character	Individual	Personality	Trait	NA	Assesses character strengths that map onto six-character virtues: wisdom & knowledge, courage, humanity, justice, temperance, and transcendence.
4	Unit Identifier	Unit	Attribute	State	NA	Unit Soldier was a member of during data collection.
5	Theft	Individual	Behavioral	Performance	Counterproductive	Act of stealing.

Note. NA = not applicable.

Table 7

Methodological Profiling of Example Metadata

Example ID#	Table Name	Data Source	Item Stem	Item Text	Item#	Data Type	Response Format	Response Values	Source Scale	Citation
1	Master	DMDC	NA	NA	NA	Administrative	Free Response	Event Date	NA	NA
2	Entry	MEPCOM	NA	NA	NA	Administrative	Categorical	2, 3, 4	NA	NA
3	GAT 1.0	ARD	Think about how you have acted in actual situations during the past four weeks. Please answer only in terms of what YOU actually did. Please read carefully. Select a number from 0 to 10 according to how often you showed/used the qualities listed.	Prudence or caution.	Q47	Deigned	Bounded Rating Scale	0 (never); 1; 2; 3; 4; 5; 6; 7; 8; 9; 10 (always)	VIA-IS	Peterson (2007); Peterson & Seligman (2004)
4	URI	ARD	NA	NA	NA	Administrative	Categorical	Alphanumeric	NA	NA
5	URI	ARD	Within the past 12 months	Have you stolen or shoplifted anything?	Q36	Designed	Dichotomous	Yes; No	NA	NA

Note. DMDC = Defense Manpower Data Center; MEPCOM = Military Entrance Processing Command; ARD = Army Resilience Directorate; NA = not applicable; VIA-IS = Values in Action-Inventory of Strengths.

Table 8

Summary of Conceptual and Methodological Profiling Interrater Agreement Results

Typology	*N*	Number of Judges	Number of Categories	Fleiss' Kappa (κ)	95% CI κ [c]
Construct Identification [a]	156	5	70	.700***	[.651, .747]
Construct Referent [a]	156	5	5	.693***	[.549, .797]
Construct Form [a]	156	5	8	.561***	[.499, .621]
Construct Framework [a]	156	5	6	.410***	[.344, .468]
Performance Type [a]	156	5	6	.452***	[.279, .590]
Data Type [b]	156	5	4	.520***	[.444, .596]

Note. [a] Conceptual Profiling. [b] Methodological Profiling. [c] 95% Confidence Interval (CI) based on 10,000 bootstrapped samples. *N* = sample size or variables judged. General interpretive guidelines for Fleiss' Kappa (Fleiss et al., 2003): .00–.40 = poor agreement beyond chance; .40–.75 = fair to good agreement beyond chance; > .75 = excellent agreement beyond chance. * $p <$ .05; ** $p <$ .01, *** $p <$ .001.

Supplemental Information

Appendix A: Detailed Walkthrough of Profiled Examples

This appendix provides greater detail on the conceptual and methodological profiling decisions (and justification thereof) for the five example variables listed in Tables 4–6 of the main text.

Example 1

From the metadata provided within the Person-Event Data Environment (PDE), the variable name (VAR_NAME) for Example 1 was 'ADSVC_PE_DT,' which had the label 'Active Duty Service Projected End Date' under the business name (VAR_BUSNAME). The ENT_NAME tells us that the variable comes from the Master Table, which typically houses administrative personnel records for different time periods in a service member's career (e.g., rank, race, home of record). Both the variable description (VAR_DESCRIPTION) and variable usage (VAR_USAGE) are provided with information. The variable description column describes the variable as a date for which a service member is expected to leave service. The variable usage column tells us that before 2000, this variable was only applicable to enlisted Soldiers.

Conceptual Profiling

Construct Identification. The construct was identified as 'Date' since the variable name and description described the variable as a date of an event (i.e., the projected end of service).

Construct Referent. The referent for the construct was categorized as 'Individual' because the variable refers to dates for individual military service.

Construct Form. The form of the construct was categorized as ‘Attribute’ because the variable refers to a characteristic of the Soldier.

Construct Framework. The variable was classified as ‘Trait’ within a conceptual framework because it serves as a characteristic marker that is unlikely to change.

Performance Type. The performance type was classified as ‘NA’ as this variable did not refer to any of the components of individual work performance outlined by Koopmans et al. (2011).

Methodological Profiling

Table Name. This table was given the name ‘Master Table’ given that this data table is often referred to as the Master File.

Data Source. The PDE provided additional information in its data catalog that this data comes from the Defense Manpower Data Center or 'DMDC' and was labeled as such.

Item Stem. This variable did not contain any item or question text and was labeled ‘NA’ during profiling for ‘not applicable.’

Item Text. This variable did not contain any item or question text and was labeled ‘NA.’

Item#. This variable did not come from a survey or form with individual item numbers, so this variable was labeled 'NA.'

Data Type. This variable was classified as ‘Administrative,’ as it came from an administrative table of personnel records.

Response Format. This variable was classified as ‘Free Response,’ since a date was what was recorded.

Response Values. This variable was classified as 'Event Date,' since values pertained to a year-month-day date format.

Source Scale. This variable did not come from a previously published measure or scale and was labeled ‘NA.’

Citation. This variable did not come from a previously published measure or scale and was labeled ‘NA.’

Example 2

From the metadata provided within the PDE, the variable name (VAR_NAME) for Example 2 was 'RECORD' which had the label 'Record Status' under the business name (VAR_BUSNAME). The ENT_NAME tells us that the variable comes from the MEPCOM Table, which typically houses administrative records for Soldiers upon accession (i.e., entry) into the Army. The variable description (VAR_DESCRIPTION) does not provide more information than the variable and business name, and the variable usage (VAR_USAGE) is blank. The variable description column describes the variables as a record, but further details are not provided.

Conceptual Profiling

Construct Identification. The construct was categorized as 'Administrative' since the variable name and description identify the variable as some sort of administrative record. However, given the lack of information, a more fine-grained construct could not be identified.

Construct Referent. Categorized as 'NA' because of the lack of information.

Construct Form. Categorized as 'NA' because of the lack of information.

Construct Framework. Classified as 'NA' because of the lack of information.

Performance Type. The performance type was classified as ‘NA’ because of the lack of information.

Methodological Profiling

Table Name. This table was labeled ‘Entry Table’ given that this data table reflects data related to initial entry into the Army.

Data Source. The PDE provided additional information in its data catalog indicating that this data comes from the Military Entrance Processing Command or 'MEPCOM' and the source was labeled as such.

Item Stem. This variable did not contain any item or question text and was labeled ‘NA’ for ‘not applicable.’

Item Text. This variable did not contain any item or question text and was labeled ‘NA.’

Item#. This variable did not come from a survey or form with individual item numbers, so this variable was labeled 'NA.'

Data Type. This variable was classified as ‘Administrative’ as the variable’s business name identifies the variable as an administrative record of some sort.

Response Format. This variable was classified as ‘Categorical’ for response format because several discrete categories were found in the actual data in integer format.

Response Values. When examining the unique values found in the data for this variable, ‘2, 3, 4’ were found as response values. The values pertain to some sort of record status categories—although their corresponding meaning is undetermined without a codebook available for this variable.

Source Scale. This variable did not come from a previously published measure or scale and was labeled ‘NA.’

Citation. This variable did not come from a previously published measure or scale and was labeled ‘NA.’

Example 3

From the metadata provided within the PDE, the variable name (VAR_NAME) for Example 3 was 'Q47', which had the label 'Emotional Fitness, Character, Q47' under the business name (VAR_BUSNAME). The ENT_NAME tells us that the variable comes from the GAT 1.0 Table, a data table containing survey data from version 1 of the Global Assessment Tool, a constellation of measures examining psychosocial characteristics (e.g., depression, coping styles, work engagement). Both the variable description (VAR_DESCRIPTION) and variable usage (VAR_USAGE) include additional information. The variable description column lists the

question asked of Soldiers when they completed the survey. The variable usage column only indicates that the variable was scored in some way; no further information is provided.

Conceptual Profiling

Construct Identification. The construct was identified as ‘Character’ since the variable description identifies the variable as being one of the items assessing character from the GAT (for a review, see Vie et al., 2016).

Construct Referent. The referent for the construct was categorized as ‘Individual’ because the variable description provides wording of the measurement item using singular pronouns related to an individual (i.e., ‘you’).

Construct Form. The form of the construct was categorized as ‘Personality’ because the variable description refers to a general psychosocial characteristic of the person as expressed through their actions towards others.

Construct Framework. The variable was classified as 'Trait' within a conceptual framework because a person's character is typically stable over time.

Performance Type. The performance type was classified as ‘NA’ as this variable did not refer to any of the components of individual work performance outlined by Koopmans et al. (2011).

Methodological Profiling

Table Name. This table was given the name ‘GAT 1.0 Table’ given that this data table reflects data related to version 1 of the Global Assessment Tool or GAT.

Data Source. The PDE provided additional information in its data catalog that this data comes from the Army Resilience Directorate or 'ARD' and was labeled as such.

Item Stem. The variable did contain a stem for the respondent's question (i.e., 'Think about how you have acted…used the qualities listed?'), which set up the specific characteristic being highlighted and question to which the individual was expected to responde.

Item Text. The item text or specific thing being asked of the respondent was their degree of ‘Prudence or caution.’

Item#. The variable name 'Q47' references the item number '47' in the survey.

Data Type. This variable was classified as 'Designed' as the variable comes from a survey designed to assess the psychosocial characteristics of respondents.

Response Format. This variable was classified as 'Bounded Rating Scale' for response format because a rating scale shows different degrees of frequency.

Response Values. The response values were a part of an 11-point Likert scale ranging from 0 (*never*) to 10 (*always*).

Source Scale. According to published literature on the variable (see Vie et al., 2016), the variable originated from the Values in Action-Inventory of Strengths scale or VIA-IS.

Citation. According to published literature on the variable (see Vie et al., 2016), the citation for the original generation of this variable or item came from several sources (see Peterson, 2007; Peterson & Seligman, 2004).

Example 4

From the metadata provided within the PDE, the variable name (VAR_NAME) for Example 4 was 'UIC_PDE,' which had the label '[PDE] Unit Identification Code' under the business name (VAR_BUSNAME). The ENT_NAME tells us that the variable comes from the URI Table, which houses survey data from the Unit Risk Inventory related to undesirable behaviors of Soldiers in units (e.g., substance abuse, crime). A variable description (VAR_DESCRIPTION) is provided for the variable but not a variable usage (VAR_USAGE). The variable description column indicates the variable is an identification code for the current unit to which a service member has been assigned.

Conceptual Profiling

Construct Identification. The construct was identified as 'Unit Identifier' since the variable description identifies the variable as a unit ID code for a Soldier.

Construct Referent. The referent for the construct was categorized as 'Unit' because the variable description identifies the variable as related to a Soldier's unit.

Construct Form. The construct form was categorized as 'Attribute' because the variable refers to a characteristic of the Soldier (i.e., the unit they are a member of).

Construct Framework. The variable was classified as 'State' within a conceptual framework because the Soldier's assigned unit frequently changes over a Soldier's career.

Performance Type. The performance type was classified as 'NA' as this variable did not refer to any of the components of individual work performance outlined by Koopmans et al. (2011).

Methodological Profiling

Table Name. This table was given the name 'URI Table' given that this data table reflects data related to the Unit Risk Inventory survey.

Data Source. The PDE provided additional information in its data catalog that this data comes from the Army Resilience Directorate or 'ARD' and was labeled as such.

Item Stem. This variable did not contain any item or question text and was labeled ‘NA.’

Item Text. This variable did not contain any item or question text and was labeled ‘NA.’

Item#. This variable did not come from a survey or form with individual item numbers, so this variable was labeled 'NA.'

Data Type. This variable was classified as ‘Administrative’ as the variable refers to an administrative record or characteristic of the Soldier.

Response Format. This variable was classified as ‘Categorical’ for response format since a code is given as a unit identifier.

Response Values. The response values were categorized as ‘Alphanumeric’ as there were numerous combinations of letter/number codes for different Army units.

Source Scale. This variable did not come from a previously published measure or scale and was labeled ‘NA.’

Citation. This variable did not come from a previously published measure or scale and was labeled ‘NA.’

Example 5

From the metadata provided within the PDE, the variable name (VAR_NAME) for Example 5 was 'Q36', which had the label 'Criminal History, Q36' under the business name (VAR_BUSNAME). The ENT_NAME tells us that the variable comes from the URI Table, which houses survey data from the Unit Risk Inventory related to undesirable behaviors of Soldiers in units (e.g., substance abuse, crime). Both the variable description (VAR_DESCRIPTION) and variable usage (VAR_USAGE) are provided. The variable description column describes the variable in terms of the question asked of Soldiers when they completed the survey. The variable usage column only tells that the variable was scored in some way; no further information is provided.

Conceptual Profiling

Construct Identification. The construct was identified as 'Theft' since the variable description identifies the variable as indicating whether Soldiers have stolen items in the past.

Construct Referent. The referent for the construct was categorized as ‘Individual’ because the variable description provides wording of the measurement item using singular pronouns related to an individual (i.e., ‘you’).

Construct Form. The construct form was categorized as 'Behavioral' because the variable refers to an act that the Soldier may have committed in the past.

Construct Framework. The variable was classified as 'Performance' within a conceptual framework because stealing or theft would fall under the counterproductive performance criteria outlined by Koopmans et al. (2016).

Performance Type. The performance type was classified as ‘Counterproductive’ because stealing or theft would fall under the counterproductive performance criteria outlined by Koopmans et al. (2016).

Methodological Profiling

Table Name. This table was given the name ‘URI Table’ given that this data table reflects data related to the Unit Risk Inventory survey.

Data Source. The PDE provided additional information in its data catalog that this data comes from the Army Resilience Directorate or 'ARD' and was labeled as such.

Item Stem. The variable contained a stem for the respondent's question (i.e., 'Within the past 12 months…'), which set up the specific statement to which the individual was expected to respond.

Item Text. The item text or specific thing being asked of the respondent was whether they had ‘stolen or shoplifted anything.’

Item#. The variable name 'Q36' references the item is number '36' in the survey.

Data Type. This variable was classified as 'Designed' as the variable refers to a survey examining risky behavior in Army units.

Response Format. This variable was classified as ‘Dichotomous’ for response format because only two categories could be chosen as a response.

Response Values. The response values were either ‘Yes’ or ‘No.’

Source Scale. This variable did not come from a previously published measure or scale and was labeled ‘NA.’

Citation. This variable did not come from a previously published measure or scale and was labeled ‘NA.’

Appendix B: Determining Interrater Agreement Example

Please find a completed excel workbook along with this supplemental material showing how interrater agreement was determined. R code is also provided for calculating interrater agreement indices.

SI References

Koopmans, L., Bernaards, C. M., Hildebrandt, V. H., Schaufeli, W. B., de Vet, H. C. W., & van der Beek, A. J. (2011). Conceptual frameworks of individual work performance: A systematic review. *Journal of Occupational and Environmental Medicine*, *53*, 856–866. https://doi.org/10.1097/JOM.0b013e318226a763

Peterson, C. (2007). *Brief Strengths Test*. Cincinnati: VIS Institute.

Peterson, C., & Seligman, M. E. P. (2004). *Character Strengths and Virtues: A Handbook and Classification*. New York: Oxford University Press/Washington, DC: American Psychological Association.

Vie, L. L., Scheier, L. M., Lester, P. B., Seligman, M. E. P. (2016). Initial validation of the U.S. Army Global Assessment Tool. *Military Psychology*, *28*, 468–487. https://doi.org/10.1037/mil0000141