
Yes Virginia, There Is An Objective Reality In Job Analysis

Robert J. Harvey
Virginia Polytechnic Institute & State University

Mark A. Wilson
North Carolina State University

We fundamentally disagree with Sanchez and Levine (in press) on several issues. Terminologically, we are troubled by their failure to differentiate between the descriptive process of rating verifiable work characteristics (i.e., job analysis) versus the subjective process of inferring worker ability and “other” (AO) requirements (i.e., job specification). Although “consequential validity” is crucial for evaluating job specifications, it is largely irrelevant for assessing properly conducted job analyses. Ontologically, we reject their relativist view that an objective reality does not exist when describing work activities. When verifiable descriptors are rated using sound rating scales, independent judges can definitively assess position rating accuracy; such a review constitutes all the “validity” evidence needed for the job analysis per se. We discuss a number of additional concerns, including the way in which practitioners deal with true cross-position ratings variability, and the role of holistic inferences.

Sanchez and Levine (in press) raised a number of important issues with respect to the ways in which researchers and practitioners have attempted to quantify the accuracy, validity, and reliability of job ratings; let us begin by stressing that we have no qualms regarding their calls for an increased focus on the consequences of using job analysis (JA) data to make personnel decisions. Few (if any) people perform job analyses simply to experience the intrinsic joy of doing so: instead, JA ratings are typically collected as a means to achieve one or more organizationally valued ends (e.g., developing a selection test battery, constructing a situational judgment test, developing an assessment center exercise, designing a training program, administering a compensation system, identifying job families, developing a performance appraisal system). As such, the practical utility and legal defensibility of these end-products represent paramount concerns – in particular, the degree to which data can be offered to justify the inferences

that were made during the course of developing the personnel systems in question. Unfortunately, in view of the numerous cases in which organizations have fallen short when challenged in court to offer evidence of the validity and utility of their personnel systems (e.g., Guttman, 1993), it is clear that Sanchez and Levine’s calls to place greater emphasis on developing systems with documented utility and validity are well taken.

Having said that, however, we must take issue with many of the points raised by these authors, ranging from relatively dry terminological issues (particularly the definition of the term job analysis) to the metaphysical and ontological assumptions that underlie their arguments. We will first provide an overview of the basic principles, terminology, and assumptions guiding our critique (Figures 1-3, Table 1), followed by a point-by-point discussion of our areas of disagreement and agreement with Sanchez and Levine (in press), including related

topics that go beyond the specific issues raised in their paper.

Basic Principles and Background

Figure 1 depicts the two “domains of human behavioral taxonomies” (e.g., Dunnette, 1976) that lie at the heart of all discussions of methods for linking job analysis ratings (the domain on the right side of the figure) with the human ability/skill constructs presumed to be necessary for successful job performance (domain on the left). Illustrative examples are provided in Figure 1 of the different types of content that might be found at varying levels of abstraction (the vertical dimension) in both domains.

Figure 2 presents an updated view of the Harvey (1991) taxonomy of JA methodologies for collecting job analysis ratings, simplified to reduce the domain of methods to four basic quadrants (denoted Type I through Type IV). As in the taxonomy presented in Harvey (1991, pp. 81-85), the critical point is that methods of collecting job analysis data are defined in terms of the combination of two continua: (a) the type of work-descriptor items being rated (the vertical axis in Figure 2), which ranges from extremely molecular (bottom) through highly abstract and holistic considerations of the job as an undifferentiated entity (top); and (b) the type of rating scale used to rate the work-descriptor items (horizontal axis), which can range from highly verifiable scales anchored in terms of verifiable, job-relevant anchors that retain a constant meaning across diverse jobs (left) through highly within-job relativistic, non-verifiable, subjective rating scales that are anchored using non-job-related content (right).

The range of possible values for the vertical axis in Figure 2 can be seen in the rightmost portion of Figure 1. For example, traditional task-based JA methods would focus on the types of items contained in the bottom row of Figure 1 (e.g., “fire warning shots at fleeing felons in vehicles using handgun”). The salient characteristic of such items is their behavioral or technological specificity, which (when properly written, and combined with a suitable rating scale) allows for high comprehensibility on the part of raters, as well as easy verifiability by sufficiently job-knowledgeable independent

reviewers. In sharp contrast, holistic JA items are characterized by their high behavioral/technological abstraction and heterogeneity (and associated lower comprehensibility and potential for verification); for example, a single item might be used to rate the degree to which “Mechanical Activities” characterize a job, or the complexity of the “Things” (in the sense of the Data-People-Things taxonomy in the DOT) function of the job.

Thus, with respect to the vertical axis in Figure 2, the critical issue concerns the degree of behavioral/technological abstraction embodied in the job descriptor item to be rated; in contrast, and largely independent of that choice, the horizontal axis varies in terms of the type of judgment that is required of the rater when he/she describes the job characteristic in question (i.e., the type of rating scale). Scales that lie at the leftmost end of this axis tend to be behaviorally specific, easily verifiable, and capable of allowing meaningful level-based comparisons across jobs (i.e., the “cross job relative” type described in Harvey, 1991, pp. 87-89). For example, the “fire warning shots at fleeing felons” item described above could be rated dichotomously in terms of whether it may ever be required on the job (a Do-You-Perform scale), or in terms of how many times it is performed in the course of a year, or on a multipoint frequency rating scale (e.g., 1 = constantly to hourly; 2 = every few hours to daily; 3 = every few days to weekly; 4 = every few weeks to monthly, etc.). When combined with behaviorally specific JA items, the resulting ratings are straightforward to collect and verify, and they allow users of the information to make meaningful, level sensitive comparisons between different jobs regarding the item in question (i.e., a given rating retains a constant meaning regardless of the job being rated).

In contrast, rating scales that fall toward the rightmost end of the horizontal dimension in Figure 2 suffer from one or more flaws (either intrinsic subjectivity, or the within-job-relative problems described by Harvey, 1991, pp. 90-91) that limit the degree to which one can collect verifiably accurate ratings or make meaningful, level-sensitive comparisons between different jobs. For example, an item such as “negotiate

with others in order to reach an agreement” could be described using the following scale: “Rate the relative importance of this activity (in relation to all of the other activities performed on your job): 1 = very low, 2 = low, 3 = average, 4 = high, 5 = extreme.” Similarly, raters could be asked to “Rate the difficulty you experience when performing this activity: 1 = easy, 2 = average, 3 = hard.” Unfortunately, judgments made with the former type of explicitly “relative” scale will tend to have an ipsative character (i.e., because they are self-normed relative to the “average” activities performed on that job, and because they lack anchors that retain an unambiguous, absolute meaning across jobs); as a consequence, they make the task of drawing cross-job comparisons more difficult, if not downright misleading. Likewise, judgments made with the latter type of scale tend to have a much higher degree of subjectivity and lower verifiability than ratings made using more concrete scales.

For example, a Head Clerk and CEO may both give a Relative Importance rating of ‘5’ and a Difficulty rating of ‘3’ to an item such as “negotiate with others” based on their judgments that, relative to the other things they do, it is of “extreme importance” to their respective jobs; the Head Clerk bases this on the fact that he or she spends five minutes a day informally negotiating with the other Clerks to schedule their work hours for the next day (an important matter from their perspective), whereas the CEO gives the rating due to the fact that he or she spends many hours per week negotiating sales agreements worth millions with an array of domestic and foreign corporations (a very important matter from the perspective of the corporation’s stockholders). Clearly, despite their identical ratings, it is highly unlikely that the negotiating activities performed by the Head Clerk and CEO are identically important to the organization, or that the task of negotiating the working schedule for an office workgroup has the same level of difficulty as the task of negotiating international sales contracts that affect the profitability of the entire organization (plus the fact that it is the worker’s perception of difficulty that is rated, which may ultimately be much more of a person-specific than job-specific quantity). Additionally, given their increased subjectivity, such ratings are much

more difficult to verify independently than ratings made with scales that lie toward the leftmost end of this continuum.

As we discuss in more detail below, in our view only the Type III methods of rating jobs are effectively consistent with the definitions and objectives set forth by Harvey (1991, pp. 74-75) with respect to the issue of collecting high-quality, defensible JA ratings. By way of clarification, we must stress that abstract types of job analysis data – for example, the job dimensions produced by combining a number of moderate-specificity ratings in so-called worker-oriented job analysis instruments such as the Common-Metric Questionnaire (CMQ; Harvey, 1991) and Position Analysis Questionnaire (PAQ; McCormick, Jeanneret, & Mecham, 1972) – unquestionably represent a highly desirable level of analysis for describing jobs; however, such data are typically derived from more elemental (and verifiable) JA ratings, not directly or holistically rated themselves. Thus, the Type I and II quadrants in Figure 2 refer to methods of rating jobs that directly attempt to judge highly abstract descriptors of work, not to the job dimension scores that can be derived empirically through combination of multiple Type III ratings (e.g., the lower two rows of information in Figure 1 on the Work Activity side).

A notion that runs throughout Sanchez and Levine (in press) and Levine and Sanchez (1999) is that that accuracy and validity are fundamentally independent concepts. Regarding validity, our position is that validation essentially involves answering two questions: (a) what inferences are being drawn (e.g., because a job involves skilled mechanical activities and tool use, that a Mechanical Reasoning test will be predictive of successful job performance), and (b) are those inferences justifiably correct (i.e., supported by data)? Although it is obvious that the specific inferences made when using job analysis data to make personnel decisions will vary as a function of the purpose being addressed (e.g., using task-level data to infer a specific KS-based requirement involves a much smaller inferential leap than using abstract job dimension data to rationally infer AO-based minimum competencies), the basic relation between job analysis accuracy versus personnel decision

validity (i.e., the ways in which JA data are used to make selection, compensation, training, etc., decisions) – which is graphically depicted in Figure 3 – should be relatively constant across the various personnel functions.

As is shown in Figure 3, we identify four general types of outcomes that are possible when using job analysis data to drive a personnel decision: (a) when the job analysis data are accurate, the inferences being drawn from those data may be (to varying degrees) valid or not valid; and (b) when the job analysis data are incorrect, erroneous, or otherwise inaccurate, the inferences drawn from those data may again be (to varying degrees) invalid or valid. However, presumably this latter possibility (the “???” quadrant, in which inaccurate JA ratings lead to a valid personnel decision) would be unusual in practice, perhaps reflecting the action of chance or dumb luck. The remaining outcomes reflect what we would rationally expect to occur: namely, accurate JA data, combined with a sound process of linking JA ratings to personnel decisions, produces the desired result, whereas inaccurate JA data and/or a flawed process of linking JA data to personnel decisions would be expected to produce invalid results (other than by chance).

Finally, a comprehensive discussion of JA accuracy and validity cannot occur unless we deal with level-of-analysis issues, aggregation of data, and the role of independent, objective review of JA ratings (topics that received little attention in the Sanchez and Levine article). That is, we must consider (a) whether the ratings are being evaluated at the level of analysis of the position (i.e., the work performed by one job incumbent) versus the job (i.e., the aggregate profile formed by combining ratings given to a number of different positions holding a common job title); (b) whether each position is only rated once, as by its incumbent or a single job analyst or subject-matter expert (SME), or whether multiple sources of ratings for each position are obtained (e.g., by a team of job analysts, or incumbent-supervisor dyads); as well as (c) the type of review or verification process that is employed (if any). Table 1 summarizes the various combinations of these factors and our evaluations of each.

A key issue in this regard is the phenomenon of aggregation bias (e.g., James, 1982), which describes the tendency of aggregate-level (e.g., group mean) profiles to become increasingly misleading – and an increasingly poor descriptor of any given member of the group being aggregated – as a function of increased within-group heterogeneity with respect to the items that make up the profile. The issue of aggregation bias is relevant at both levels-of-analysis presented in Table 1 (i.e., forming a position-level aggregate profile by collapsing across multiple judges, and forming a job-level aggregate by collapsing across position profiles), and it is of critical importance in job analysis for two main reasons: (a) when nontrivial cross-judge or cross-position disagreement in the JA items is present, the aggregate profile presents an increasingly misleading summary of the activities performed (and the levels at which they are performed) by the “typical” position incumbent (e.g., the group-mean profile may not accurately describe the activities of any of the group members when high cross-position variability is present); and (b) correlations comparing the profiles of individual raters, positions, or jobs with the aggregate profile (a common, but questionable, practice in studies evaluating

holistic ratings of ability trait requirements; e.g., Fleishman & Mumford, 1991) can present a highly inaccurate index of cross-rater or cross-group agreement as the level of within-group variability increases (e.g., Harvey & Hayes, 1986; Harvey & Wilson, 1998).

Specific Points of Contention

With the above discussion of general issues serving as background, we proceed to a point-by-point critique of the Sanchez and Levine (in press) arguments. As we noted earlier, we do not take issue with every conclusion offered by these authors; in particular, we echo their call for increased efforts to validate the ways in which JA data are used to solve organizational problems and make personnel decisions. However, many basic points of disagreement remain; the first two points below focus on our most fundamental differences, whereas the remaining points address areas of either partial agreement, or topics that are related to these big-picture issues but not directly assessed by Sanchez and Levine (in press).

1. Job analysis and job/worker specification represent fundamentally different processes that should not be given a shared name. One of our most basic disagreements with the Sanchez and Levine (in press) arguments concerns terminology: namely, their practice of using the term “job analysis” to refer to virtually any activity that involves making ratings regarding work activities or the prerequisite worker-traits needed for successful job performance, regardless of the amount of subjectivity. For example, “job analysis ...is the process of gathering, analyzing, and structuring information about a job’s components, characteristics, and job requirements” (Sanchez & Levine, 1999, p. 3, emphasis added); likewise, “JA is essentially a tool intended to facilitate inferences involving job requirements” (p. 12, emphasis added).

In response, we must stress that the definition of JA adopted by Sanchez and Levine (in press) – although certainly not without precedent (e.g., Fleishman & Reilly, 1992) – is fundamentally at odds with the longstanding practice of clearly distinguishing between a job analysis (i.e., the process of describing what is done on a job, and the context in which the work activities are

performed) versus a job specification, or JS (i.e., the process of inferring the human-trait requirements presumed to be necessary for successful job performance). Historically, this distinction has been made quite clearly (see Harvey, 1991, pp. 74-78 for a review) in both the research literature as well as professional and governmental guidelines that pertain to the employee selection and assessment process (e.g., APA Standards, 1985; Uniform Guidelines, 1978). This fundamental distinction between JA versus JS was reflected in the definition of JA offered by Harvey (1991, p. 74), which explicitly excluded speculations regarding “job requirements” that are couched in terms of inferred worker knowledge/skill (KS) requirements and hypothetical ability or “other” (AO) traits.

Unfortunately, some authors persist in choosing to blur this distinction by using the term “job analysis” to refer to both the description of work, as well as the process of inferring worker-trait requirements, “competencies,” and similar speculative judgments (e.g., Fleishman & Mumford, 1991; Hughes & Prien, 1988; Lopez, Kesselman, & Lopez, 1981). Numerous reasons exist for arguing against this practice (see Harvey, 1991, pp. 75-78 for a more detailed discussion), including the types and magnitudes of inferences required, the cognitive decision making processes involved, the amounts and levels of specificity of information that must be considered, and perhaps most important, the fundamental differences in the types of properties being rated (i.e., properties of jobs, expressed in terms of work behaviors and contextual characteristics, in the case of job analysis, versus properties of people, expressed in terms of hypothetical psychological traits, in the case of job specification). This basic distinction regarding the object of the rating process is graphically illustrated in Figure 1; although in some cases an obvious similarity exists between the kinds of attributes being described (e.g., Skilled Mechanical Activities versus Mechanical Ability), the critical point is that a JA describes what is to be accomplished on the job or how the work is to be performed, whereas a JS engages in speculation regarding the personal traits that might be needed to successfully perform the job.

Clearly, these are fundamentally different.

It has been suggested by a number of authors that the term “job analysis” should be replaced by a new term; for example, Levine & Sanchez (in press) proposed renaming it work analysis (WA), and various others have effectively renamed JA (and/or JS) as “competency modeling.” Without going into the merits of changing the label applied to job analysis – which might not be a bad idea, given the image problem that JA faces in many quarters (e.g., Cunningham, 1989) – in our assessment it is critical that we give the process of describing work activities (whatever we choose to call it) a different symbolic label than that given to the process of inferring worker-trait requirements phrased in terms of AO-based constructs.

The relevance of this terminological issue to the Sanchez and Levine (in press) critique is straightforward: had they directed their “consequential validity” arguments solely toward the speculative inferences involved in the JS or “competency” inference/rating process (as well as the inadvisable Types I, II, and IV procedures for JA shown in Figure 2), we would be in total agreement. Clearly, the validity of speculative worker-trait inferences is entirely use-dependent, and in need of empirical “consequential” validation evidence (e.g., a criterion-related validation study showing that the identified AO traits are truly predictive of job success). JS speculations cannot simply be assumed to be valid because they were made by “experts,” or because high interrater agreement statistics were obtained when multiple judges made the AO-trait ratings.

Unfortunately, Sanchez and Levine’s definition of job analysis makes no distinction between inherently verifiable Type III JA ratings versus either speculative JS inferences, or subjective JA methods (indeed, it seems to place much more emphasis on the speculative JS-type inferences over traditional descriptive JA activities), and as a consequence we must firmly reject their contention that results-based methods must be used when assessing the quality or accuracy of true JA ratings. That is, in view of their inherently descriptive, verifiable nature, we contend that Type III JA ratings do not require the sort of use-dependent, “consequential

validation” presumed to be necessary by Sanchez and Levine (in press).

In essence, “consequential validation” is an issue of documenting the correctness of the inferences involved in using or applying JA data to make a personnel decision (for example, to infer worker-trait based job specifications or “competency” requirements). As Figure 3 illustrates, this issue is logically quite separate from the issue of assessing the accuracy and correctness of the JA ratings themselves (which, as we discuss in more detail below, can indeed be determined without any need to consider the potential uses to which the JA ratings may eventually be put). Thus, rather than focusing on “consequential validity,” in our assessment the “validity” issues that are of concern when evaluating JA ratings involve (a) the accuracy of the individual position ratings (considered in detail next), and (b) the amount of true cross-position variability that is present in a given job (i.e., as true within-title heterogeneity increases, it becomes increasingly misleading and invalid to attempt to interpret the aggregate view of the job as providing an accurate reflection of what is truly done on that job).

2. There can indeed be an “objective reality” and assessment of accuracy in JA. Moving to the metaphysical issue of ontology in the world of work, we next consider Sanchez and Levine’s views that (a) we must reject the notion that “there is some underlying ‘gold standard’ or unquestionably correct depiction of the job” (p. 4), (b) “accuracy may be relative but not absolute” (p. 6), (c) “[rating] errors in classical reliability theory do not need to be mistakes” (p. 7), and (d) “the dubious utility of proxy true scores” (p. 10) should cause us to abandon any definition of JA accuracy that involves deviations from a standard. Although it is comforting to note that Sanchez and Levine grant that they “do not wish to argue that accuracy in JA does not matter” at all (p. 16), the gist of their argument remains quite clear: namely, that it is effectively impossible to directly document or assess the accuracy of the JA database itself; instead, we should simply concern ourselves with inferring that the JA ratings are accurate by virtue of finding “consequential validity” in the ways in which the JA ratings are subsequently applied to

solve personnel problems and make decisions.

We must confess to being puzzled by Sanchez and Levine's arguments, particularly their clearly articulated conclusion that it is impossible to assess the accuracy of JA ratings in the abstract (i.e., without reference to some bottom-line personnel decision that was, to some greater or lesser degree, influenced by or based on the JA ratings). Somewhat analogous to the assertion that a tree that falls alone in the forest makes no sound, this argument quite clearly implies that a job analysis database that has not yet been used to develop an end-user personnel function (e.g., selection test requirement, performance appraisal form, training program) cannot be evaluated with respect to whether or not it provides an accurate reflection of each job's true activities. As they put it, JA "accuracy does not exist in a vacuum but is defined precisely by the consequences of [using] such data" (p. 16).

In view of the many possible levels-of-analysis and modes of data collection summarized in Table 1 and Figures 1-2, we can identify only two general types of circumstances in which we might agree with Sanchez and Levine's conclusion that a "gold standard" (p. 4) of JA accuracy is impossible: namely, if we (a) limit our consideration to only the subset of highly questionable JA data-collection methods (e.g., using raters who have limited job knowledge to holistically rate vague descriptors using poorly-anchored, single-item rating scales); or (b) adopt a radical relativist ontological view that maintains that there is essentially no "reality" in the world of work, and that no matter how hard we try or how sophisticated our data-collection technologies become, job behavior simply cannot be subjected to objective analysis or verification.

This latter position is analogous to the Schmidt, Hunter, and Pearlman (1981) critique of behaviorally oriented job analysis (summarized by Harvey, 1991, pp. 101-104), which attacked task- and behavior-based JA on the grounds that work behaviors cannot be directly observed or rated. The basis for this claim was their observation that "the field of personnel psychology came to be so far off base" because during the "late 1950s and early 1960s ... behaviorist influences began to make themselves felt" (Schmidt et al., 1981, pp. 178-179, emphasis

added). In particular, they claimed that it was futile to attempt to rate work behavior or environmental characteristics, claiming that

"there are two central claims to the modern behavioristic beliefs as they are manifested in the field of personnel psychology: that [human] abilities are not observable and that behavior is observable. Both claims are false to fact... Consider the supposed observability of behavior. Suppose that a worker is to screw a certain bolt into a certain hole in each automobile as it passes on the assembly line. Is 'screwing in the bolt' an observed behavior in the worker? Certainly not." (p.181, emphasis added).

For reasons summarized in Harvey (1991, p. 104), we see little justification or benefit in adopting a metaphysical position that holds that reality is an illusion, or an entirely socially constructed artifact, especially if one is trying to do research or practice in such an inherently descriptive and "real world" domain as job analysis! In essence, adopting the ontological view that behavior and the work environment cannot be objectively rated by an observer or job analyst – or, conversely, that work behavior is as observable as human ability trait requirements (Schmidt et al., 1981, p. 181) – is tantamount to saying that the basic objective of job analysis (i.e., the description of work activities and contextual characteristics) is unachievable (or alternatively, that it is possible to directly and accurately rate unobservable, hypothetical, psychological constructs and traits using holistic rating scales).

However, Sanchez and Levine (in press) appear to rely on just this type of relativist viewpoint to form the metaphysical foundation for their arguments; for example, "in JA, just like between observers sitting on opposite sides of the Pyrenees, accuracy may be relative not absolute" (p. 6). Likewise, "JA accuracy should be defined by the consequences of the data and that, given the lack of 'true score' criteria in field JA, definitions of accuracy that do not observe this

premise have little utility” (p. 9); and JA ratings “disagreement may simply indicate systematic depictions of alternate but equally valid views” (pp. 7-8).

When considering the Type III JA methods shown in Figure 2 (which, in our assessment, are the best ones to use when collecting JA ratings), we flatly reject this assessment. By way of background, it is important to recall (see Figures 1-2) that job behaviors and characteristics of the work environment vary in terms of their observability and independent verifiability across a continuum, ranging from quite easily observable and verifiable (e.g., well-written task and duty statements, and many “worker oriented” items) at the high- and moderate specificity ranges of the continuum, to decidedly not-easily-observable at the high-abstraction end of the scale (e.g., job dimensions). Although those who subscribe to the Schmidt-Hunter-Pearlman antagonism toward “behaviorist influences,” as well as those who endorse the notion that behavior cannot be directly observed might well disagree with us, we think that it is self-evident that given proper rating conditions – especially, motivation to rate accurately, and adequate rating scales – the kinds of work activities illustrated at the bottom of Figure 1 can be described with near-perfect accuracy by SMEs who are sufficiently familiar with the position in question (of course, we do not mean to imply that unverified position-incumbent ratings should simply be presumed to be accurate, but rather that the potential to achieve a “gold standard” of accuracy does indeed exist; see Point 3 below for further discussion).

By way of example, we can state with 100% confidence and accuracy that none of the four tasks listed in the bottom row of work activities in Figure 1 are performed on either of the authors’ positions. We would likewise hope that if they were called upon to rate their own positions using these items, Drs. Sanchez and Levine would be similarly able to unambiguously and accurately tell us if firing handguns at fleeing felons in vehicles, making measurements using a dial-indicator or Swiss-hole micrometer, or reprogramming tape-controlled milling machines are – or are not – activities that occur when they perform their positions. When considering JA

ratings of this nature, we find it impossible to accept the Sanchez and Levine argument that an objective reality does not exist, that a “gold standard” of accuracy cannot be determined if one is willing to exert the effort required to find it, or that JA accuracy is only a “relative” concept that can never be definitively established in the abstract (i.e., without reference to the validity of a subsequent use of the JA ratings).

The other potential situation in which we could agree with Sanchez and Levine’s (in press) rejection of the notion of an objective reality or the possibility of obtaining an “unquestionably correct depiction of the job” (p. 4) would result if we were to consider only the subset of undesirable modalities for collecting JA data presented in Table 1 and Figures 1-2 (or, if we were to include speculative JS inferences under the label of “job analysis,” which we do not). For example, although we strongly advise against doing so, it is not uncommon for JA practitioners to do things such as (a) collect data from multiple positions in each job using only the job incumbents as ratings source, (b) fail to subject the position ratings to independent review by qualified and unbiased SMEs to verify their accuracy and correct inaccuracies, (c) fail to resolve cross-rater disagreement when multiple judges rate a common position prior to forming a position-level aggregate profile, (d) use a JA methodology that is so imprecise or abstract that it is impossible to objectively review and verify the ratings, even by a highly job-knowledgeable SME (e.g., by using “relativistic” rating scales, rating scales with imprecise – or missing – anchors, or highly abstract, holistic work-activity items), and (e) eliminate cross-position and cross-rater disagreement by computing a mean profile for each job, based on the logic that “individual biases should cancel each other out according to the axioms of classical reliability theory” (Sanchez & Levine, 1999, p. 11). In such cases, we will gladly grant Sanchez and Levine’s assertion that it is futile to hope to find an “unquestionably correct depiction of the job” due to the numerous, ill-advised decisions that were made.

Simply put, the use of poor JA methodologies (in particular, using ambiguous, poorly anchored, or within-job-relativistic rating scales, as well as

attempting to rate highly abstract, unobservable, or multidimensional work attributes) will lead to a situation in which independent verification of the JA ratings is next to impossible, due to the inherent lack of objectivity of the rating items and/or scales. There is simply no getting around the fact that the Type I, II, and IV methods of collecting JA data shown in Figure 2 involve – in our assessment – an unacceptably high degree of subjectivity and speculation on the part of the raters, no matter how well motivated or job-knowledgeable they may be. Such ratings cannot be “validated” via cross-judge agreement or “interrater reliability” due to their fundamentally speculative nature; instead, they must be empirically validated in the same fashion as any other speculative judgment.

It is critical to stress that JA practitioners are not restricted to using hopelessly subjective strategies for collecting job analysis ratings; indeed, although we cannot cite precise figures to document the percentage of studies using the four types of methods shown in Figure 2, it is our assessment that many – if not most – JA projects are of the desirable Type III variety (i.e., studies in which reasonably observable aspects of work are rated using scales that possess reasonable psychometric characteristics). For example, in the popular task-inventory approach to job analysis (e.g., Christal, 1974), items such as “Use Swiss-hole micrometer to adjust bottling machine,” “Sever jugular vein of animals or poultry to be slaughtered,” “Compute variation in lengths of wire to obtain specified resistance,” or “Heat mechanical components in furnace to sinter fire-metallized coatings” might be rated on a dichotomous “do you perform” scale or a multipoint frequency or time-spent scale. Likewise, in a worker-oriented instrument like the CMQ (Harvey, 1991), items such as “Contact technical specialists who have supervisory responsibilities for the purpose of formally bargaining or negotiating,” “Summarize or condense written words in a foreign language,” or “Assemble, disassemble, or repair electrically powered tools” would be rated on applicability and frequency scales.

In such situations, we find it impossible to agree with Sanchez and Levine’s conclusion that “the accuracy of the JA data is defined by their

effects, and it cannot be determined in the absence of knowledge of such effects” (p. 10, emphasis added). On the contrary, we maintain that the accuracy of Type III JA ratings can indeed be verified – quite easily, in fact – via a review of the position ratings by independent, job-knowledgeable SMEs. In our view, this accuracy review constitutes all of the “validation” evidence that is either necessary or appropriate for a position-level JA database. In particular, it is not necessary to determine the “consequential validity” of the JA ratings at this point, because said “consequential validity” would unavoidably represent a property of the subsequent uses of the JA data to make specified personnel decisions (e.g., to identify a selection test battery and set cutoff scores) and the decision processes that were used when linking the JA data to the subsequent personnel decision, which is clearly not a property of the JA database.

In our assessment, it is self-evident that the accuracy and correctness of Type III JA ratings can be assessed without the need to specify even a single potential use of the JA data. Using the above examples, the reality of whether or not a worker is required to shoot guns at fleeing felons, use Swiss-hole micrometers, or slit the jugular veins of poultry being slaughtered is quite invariant of the eventual uses of such ratings: they either do, or they do not, and whether they do or do not is in no way dependent on the desired use of the JA ratings. Likewise, the degree of “consequential validity” seen in one potential use of the JA data – for example, using task ratings to develop a work sample test – obviously has no necessary relation to the validity of using the same data to drive a different purpose (e.g., developing minimum “competency” requirements for the job). Indeed, under Sanchez and Levine’s view that the quality of the JA database can only be defined by the use of that information to make a personnel decision, it would appear to be easy to produce internally contradictory or inconsistent conclusions regarding the accuracy of the JA database. For example, if one finds that the JA ratings lead to valid decisions when used to develop one application (e.g., a work-sample), yet they lead to invalid decisions when used to develop a second application (e.g., setting “competency”

requirements), should one conclude that the JA database is: (a) valid, (b) accurate, (c) not valid but accurate, (d) valid but not accurate; (e) not accurate or valid, or (f) semi-valid?

We maintain that the correct answer is “none of the above.” Clearly, as is diagrammed in Figure 3, the inherent quality or accuracy of the JA database is not the result of some subsequent use of it to develop or drive a specific personnel decision (indeed, to the extent that there is any causal relation present, the true direction of causality presumably flows in the opposite direction). On the contrary, position-level JA accuracy, and the degree of true cross-position variability (to be discussed below), exist independent of potential uses of such data, and it is our view that these factors should be our primary focus when collecting and evaluating JA ratings. Discussions of the validity of subsequent uses of the JA data to drive specific personnel functions that may be many, many steps removed from the JA itself must be deferred to those subsequent uses, and not viewed as “the better standard for [assessing] job analysis data” (Sanchez & Levine, p. 1).

In sum, on the topic of JA ratings accuracy – and whether or not we should even attempt to directly assess it – we reach diametrically different conclusions than those offered by Sanchez and Levine: namely, that an objective reality can indeed be defined, and that properly used Type III job analysis rating methods are quite capable of describing it. However, we must stress that the only way in which one can confidently assess JA accuracy is via independent SME review of position-level JA ratings. In our assessment, such an accuracy review constitutes the only “validation” that is necessary for the position level JA database; the “consequential validity” of any subsequent inferences based in whole or in part on the JA database – although clearly important – represents an entirely separate issue that has nothing to do with job analysis per se.

3. SME judgments made by objective job analysts are likely to be superior to ratings made by typical position incumbents, especially when real-world outcomes are at stake. At the risk of being termed cynical, we likewise disagree with Sanchez and Levine’s conclusion that “the

superiority of SME judgments should not be taken for granted, unless such judgments are proven consequential” (1999, p. 11). Simply put, in our view there is no justification for the widespread practice of blindly hoping or trusting that job incumbents will make accurate ratings when there are no contingencies in place to hold them accountable for the accuracy and correctness of those ratings, especially when the ratings are not reviewed independently for accuracy (an extreme example of this type of misplaced hope, which apparently is not uncommon in practice, occurs when incumbents make their JA ratings anonymously).

We will be the first to admit that many JA practitioners take incumbent ratings of their positions at face value, and fail to include any sort of formal SME accuracy review as part of the JA processes. However, in the absence of such a review, it should be immediately apparent that in such situations one has absolutely no assurances regarding the quality or accuracy of the JA database, and that it is likewise unacceptable to attempt to use cross-position interrater agreement as a surrogate index of JA ratings quality (a point on which we agree with Sanchez & Levine).

As we note in our Table 1 recommendations, it is our strongly held view – based on over 40 years of combined experience conducting job analyses – that it constitutes a very serious mistake to assume that position incumbents or similar raters will provide accurate, unbiased ratings. Indeed, perhaps our most important recommendation is that JA practitioners should always include a formal accuracy review using objective, properly motivated SMEs in all cases in which incumbents, supervisors, or similar raters having questionable motives and/or instrument familiarity and rating experience are used as data sources.

Several studies (e.g., Harvey, 1990; Stutzman, 1983; Wilson, Harvey, & Macy, 1990) have offered empirical support for this assessment, showing that major mistakes and inconsistencies are probably the norm – rather than the exception – when JA ratings are obtained from anyone other than a trained, job-knowledgeable, objective job analyst (and questions can even be raised about job analysts, especially when they are required to make ratings of highly abstract work activities;

e.g., Butler & Harvey, 1988). For example, Stutzman (1983) reported numerous cases in which incumbents claimed to perform tasks that were known to not be part of the job in question (a number of known-irrelevant items were deliberately included in the task inventory). In the Harvey (1990) study, task inventory ratings by incumbents were subjected to independent review for accuracy by their supervisors, and numerous – and often, egregious – errors were uncovered (e.g., police officers who claimed to drive ambulances; water-meter readers who claimed that they did not record water usage by reading gauges, janitors who claimed they recorded the official minutes of city council meetings).

In a recent project completed by the first author using a worker-oriented questionnaire, similar results were observed (e.g., parking lot cashiers claimed that they delegated work assignments to executives; nonsupervisory clerical workers claimed to have formal supervisory responsibility over other clerical employees; employees grossly exaggerated their spans of control; a records clerk claimed to have final decision-approval authority over long-term business strategy decisions; a laborer claimed to provide treatment/therapy to mid-level managers). Finally, in the context of task-inventory JA, Wilson et al. (1990) found that when a subset of tasks were repeated in a long inventory, often distressing levels of rating inconsistency were observed (in many cases, tasks that were rated does-not-apply [DNA] in one exposure were rated as being applicable the other time). Although we are sensitive to the problem of changes in JA ratings occurring over time, there is certainly no reason to assume that true changes in the work occurred over the span of time necessary to complete the JA rating survey!

As a practical matter, constructing job-related, verifiable, and comprehensive job analysis rating instruments represents only half the battle: until the ratings obtained using these instruments are subjected to independent review for accuracy, they should be presumed to contain numerous and potentially serious errors, especially when they are collected from incumbents, supervisors, and others who have a vested interest in influencing

the outcomes of the JA process. In situations in which incumbents are used as raters, it is typically the case that they are well aware of one or more intended uses of the JA ratings, and that they clearly harbor a vested interest in influencing those outcomes (e.g., when the JA is being done to develop a pay system, or set selection or promotion requirements). In such settings, there can be little question that incumbents, supervisors, and other “non professional” SMEs are often amply motivated to provide a distorted (typically, exaggerated) view of their job. Additionally, independent of an overt motivation to distort, factors such as response fatigue (especially in long instruments), lack of adequate verbal ability or reading skill, carelessness, etc., represent equally serious potential sources of inaccuracy in position ratings made by incumbents or similar raters.

In sum, placing one’s hopes on the prospect that inaccurate, exaggerated JA data will somehow be able to produce a valid, useful personnel outcome (i.e., the “???” quadrant in Figure 3) is, in our assessment, naïve in the extreme. We feel strongly that the time to worry about assessing (and fixing) JA accuracy is before one invests large amounts of time, effort, and expense in developing end-user personnel applications (e.g., a selection test battery, assessment center, work sample, etc.). Postponing any meaningful consideration of the topic of JA accuracy until after one (a) develops a personnel application using the JA ratings as guidance, (b) conducts an empirical validation study, and then (c) evaluates the findings of the validation study is, in our assessment, a highly risky and potentially quite wasteful strategy (especially if the job analysis, development work, and validation study have to be entirely redone). This would especially be the case if an objective review by unbiased SMEs is not included as a formal step in the JA process. Although much more research is needed to determine the true “base rate” and severity of JA rating errors in unreviewed JA data, the results of existing research strongly suggest that there is no basis for assuming that such data will be free from widespread, often serious, distortions and inaccuracies.

4. True cross-position variability and

aggregation bias are critical concerns. We agree with Sanchez and Levine's assertion that using cross-position agreement as an index of JA accuracy has its limitations (e.g., p. 5); however, we reject the associated implication that within-title, cross-position variability implies that there is no "objective reality" in JA, or that "equating disagreement with inaccuracy" (p. 7) is fundamentally erroneous. Clearly, there are indeed times in which true cross-position variability in JA ratings can be confounded with ratings inaccuracy (e.g., in Table 1, the Type C case in which ratings are collected in the absence of SME verification, and without multiple ratings of each position). In such a situation, it is indeed impossible to separate cross-position disagreement due to rating errors from true cross-position variability in job duties. However, this need not be the case, and we strongly advise against intentionally designing a JA study in a fashion that deliberately confounds true cross-position heterogeneity with position rating inaccuracy.

When verifiable aspects of work are being rated with sound rating scales (i.e., the Type III JA rating methods in Figure 2), if raters judging a common stimulus produce ratings that disagree (e.g., when multiple SMEs rate a given position), we contend that this by definition constitutes inaccuracy. That is, although we may not be able to prove (without further independent review) that any of the raters have provided an accurate description, nontrivial disagreement in such a situation does indeed prove that some – and perhaps all – of the raters are rating inaccurately. For this reason, we recommend that the JA process always begin with position-level ratings, and that each position profile be verified for accuracy. After erroneous position ratings are corrected, one can then unambiguously characterize all of the remaining cross-position variance as reflecting true within-job heterogeneity.

Of course, the issue of how one deals with significant true within-title variability must still be addressed; this is where the concept of aggregation bias (James, 1982) becomes central. That is, although the existence of true cross-position variability is to some degree inevitable (e.g., due to such factors as seniority,

performance level, in-group versus out-group status with the supervisor, degree of negotiating latitude), potentially serious problems occur as the levels of within-title heterogeneity increase. As James (1982) noted, attempts to meaningfully interpret the aggregate profile in such situations can be pointless and highly misleading, due to the fact that the job-mean profile becomes an increasingly meaningless descriptor of the "typical" or "average" position's work activities. It is our assessment that the practice of simply ignoring nontrivial true cross-position disagreement by eliminating it via the computation of a job-mean profile, and then hoping that personnel decisions made on the basis of this (misleading) aggregate profile will somehow demonstrate "consequential validity," is fundamentally misguided. With sufficiently high cross-position disagreement, it can easily be the case that none of the positions holding the job title are accurately described by the group-aggregate profile; in such situations, we can see no basis for believing that sound personnel decisions could ever be reached – other than by sheer luck – on the basis of the job-aggregate results.

As with our earlier consideration of whether an "objective reality" exists in JA, we suspect that much of our disagreement with Sanchez and Levine on this matter stems from the fact that their statements are based on an overly broad definition of JA (i.e., one that includes trait-based, job specification inferences), combined with a consideration of only a subset of the possible methods for collecting JA data (e.g., those in which true cross-position disagreement is confounded with position rating errors). That is, if one considers only speculative ratings of hypothetical, AO-based worker trait requirements (i.e., JS) or the undesirable Type I, II, and IV JA rating methods in Figure 2, we would agree with their assessment that cross-rater agreement (or disagreement) says little or nothing with respect to the accuracy of the job ratings.

However, when verifiable work activities are rated using adequate rating scales (i.e., Type III JA ratings in Figure 2), and one considers cross-rater disagreement at the position level of analysis (i.e., the Type B and D cases in Table 1), quite the opposite is true. In such situations, it is

simply not the case that “disagreement may simply indicate systematic depictions of alternate but equally valid views” of reality (Sanchez & Levine, 1999, pp. 6-7). In short, when multiple judges rating a common stimulus (in this case, a position) disagree, this must constitute a mistake, and the source of such disagreement must be identified and resolved during the accuracy review process. Returning again to the domain of metaphysics, it has long been recognized that the position is the only thing that is truly “real” in an organization, and that the concept of a job is an organizational convenience applied to simplify personnel administration (e.g., Harvey, 1991, p. 79). JA researchers have likewise long known (e.g., Cragun & McCormick, 1967; Pass & Robertson, 1980) that nontrivial true, cross-position heterogeneity exists. The important point is that the existence of true disagreement does not imply that an objective reality cannot be defined, or that a “gold standard” of accuracy cannot be identified; instead, it simply means that accuracy must first be assessed at the position level of analysis, and once inaccuracies have been resolved at that level, the severity of the remaining true within-title heterogeneity must be addressed.

Once true cross-position disagreement in job activities reaches a critical level, this situation must be dealt with by the organization (typically, by revising the job title system to more accurately associate job titles with actual duties and reduce true within-title heterogeneity). One of the most useful definitions of the term “job” that we have encountered holds that a job is the collection of positions that are sufficiently similar in terms of their work activities that they can meaningfully share a common job analysis (e.g., see Harvey, 1991, p. 79). Clearly, when true within-title heterogeneity becomes excessive and aggregation bias becomes operative, this definition is no longer satisfied, and changes in the job title system must be implemented

Finally, we must note that we agree with Sanchez and Levine’s call for more research to identify factors that “moderate” ratings of a given job (i.e., the reasons that true cross-position disagreement exists). However, it is important to remember that even if we eventually find consistent types of moderation (e.g., higher

performing employees rating differently than lower performing employees), this should not be interpreted to mean that accuracy of the JA ratings is no longer important, or that the existence of moderation implies that JA accuracy cannot be assessed (as seems to be the conclusion advanced by Sanchez and Levine; e.g., pp. 6-7).

5. Both the type of attribute being rated, and the type of rating scale, must be considered when determining validation/accuracy standards for JA and JS ratings. In a nutshell, our overriding concern with the Sanchez and Levine (in press) paper is that they have combined an overly broad definition of “job analysis” (i.e., which subsumes both JA and JS ratings) with a one-size-fits-all conclusion regarding appropriate procedures for assessing ratings quality (i.e., an indirect inference of “consequential validity” via consideration of subsequent decisions that apply the JA ratings, combined with a denial that it is even possible to conduct a direct assessment of accuracy). In our assessment, the question of assessing accuracy or validation standards for JA or JS ratings is much more complicated, in part because JA and JS have such fundamentally different goals, and in part because both the JA and JS rating processes involve the combination of two largely independent factors: (a) the descriptor items being rated (e.g., tasks, elements, or duties in JA, versus required worker knowledge, skill, ability, and “other” traits in JS); and (b) the rating scale used to describe or infer aspects of the descriptor in question (e.g., time-spent or do-you-perform in JA, versus the amount or level of an ability trait required for job performance in JS). Thus, although Figure 2 was designed to describe the JA rating process, the same rationale can be applied to characterize the JS rating processes as well.

The point we wish to emphasize is that compromising the quality or verifiability of either the attribute being rated (vertical axis in Figure 2) or the scale used to describe it (horizontal axis) will produce a corresponding decrement in the quality or verifiability of the resultant rating. Thus, even if highly verifiable item content is rated in a JA context (e.g., molecular task statements), using an imprecise or within-job relativistic rating scale (see Harvey, 1991, pp. 82-84, for further discussion) can transform an easily

verifiable rating process into one that is as subjective and validity-challenged as a single-item holistic rating of a hypothetical worker trait (discussed in more detail in the following section).

Casting the JS rating process in the Figure 2 rationale used to evaluate JA rating methods implies that some types of methods for making JS inferences may be much more likely to produce valid results than others. As with JA ratings, this leads us to hypothesize that Type III JS processes (i.e., in which well-defined, specific worker characteristics or KS items are rated using well-developed, minimally subjective rating scales) would be the one most likely to produce useful results. Conversely, we hypothesize that the Type II variety of JS ratings – which are analogous to the rating instruments being used to collect data for the O*Net database (i.e., rating a small number of extremely abstract AO constructs using single-item, sparsely anchored scales composed of anchors listing largely job-irrelevant activities) – would be unlikely to be capable of producing high quality results. Given their implications for practice, we conclude that such predictions need to be thoroughly assessed in future JA and JS research; for reasons discussed below, we see a much higher potential contribution for such research in comparison to highly controlled, lab studies designed to test cognitive theories of the JA or JS rating process.

6. Holistic ratings are unlikely to ever produce high-quality data. One of the most controversial issues in job analysis today concerns the role of holistic ratings, which in the present sense can be defined as Type I or Type II JA or JS ratings at the highest end of the vertical continuum (descriptor abstraction) shown in Figure 2. By way of definition, the salient characteristic of a holistic rating task – in either the JA or JS domains – is that it involves a direct attempt to rate a highly abstract, nebulous characteristic (either general work activity, or an abstract worker ability trait) using a very simple (often, single-item) judgment process. The ability rating scales of the Fleishman system, which form the basis for a portion of the online replacement for the Dictionary of Occupational Titles (i.e., the O*Net; Peterson, Mumford, Borman, Jeanneret, Fleishman, & Levin, 1997),

typify this approach in the area of JS. For example, a single 7-point BARS-type rating scale (with anchors provided for three of the seven rating points) is used to rate the AO trait of Oral Comprehension; the rating anchors that define the scale metric are 2 = “understand a television commercial,” 4 = “understand instructions for a sport,” and 5.5 = “understand a lecture on metaphysics.”

Although holistic ratings are most commonly seen in the domain of JS rating, they can also be used to make JA ratings of highly abstract types of work content (e.g., such as the ratings of “broad categorizations of primary job responsibilities” advocated by Sanchez and Levine, p. 10). For example, in a study evaluating the quality of holistic JA ratings of the job dimension scores produced from the PAQ, Butler and Harvey (1988) asked different groups of raters – including professional job analysts with extensive familiarity with the PAQ – to directly rate, using a single-item rating scale, each of the abstract PAQ job dimensions. The criterion of accuracy or convergence in this study was the profile of PAQ job dimension scores computed in the typical, “decomposed” fashion (i.e., by statistically combining the ratings of a large number of more behaviorally specific PAQ items to produce each dimension score); as hypothesized, Butler and Harvey (1988) found very poor levels of convergence (r s in the .20’s and lower) between the direct holistic ratings of the JA dimensions and the dimension scores formed from multiple ratings of more behaviorally detailed items.

Our concerns with the holistic rating approach, when used in either JA or JS settings, are twofold: (a) in view of the well-documented limitations of human judges to perform well when large amounts of information must be combined to form summary ratings (reviewed, but largely dismissed, by Sanchez & Levine, pp. 8-10), we would not expect to find adequately high convergence between direct holistic ratings and the corresponding decomposed scores (if available); and (b) when presumed-accurate decomposed scores are not available to serve as a basis for evaluation, the fact that the holistic traits are so far removed from observable, verifiable work reality makes it virtually impossible (short

of conducting a criterion-related validation study, or similarly ambitious empirical undertaking) to justify and document their accuracy (in the case of JA) or their validity (in the case of JS). Without doubt, cross-rater agreement or “interrater reliability” does not provide any proof of the accuracy or validity of such data.

Sanchez and Levine appear to endorse the notion of using holistic ratings of JA and JS characteristics, as well as “competency” requirements; for example, “one may even argue that [ratings of] the broad job responsibilities are more cost-effective than the costly and cumbersome task inventories” (p. 10). In contrast, we remain highly skeptical; in our assessment, the available (albeit limited) research suggests that it is quite unlikely that holistic ratings of either JA or JS constructs will ever be found to perform acceptably, even when high quality rating scales are used. In the area of JA, the Butler and Harvey (1988) results demonstrated that the correlations between holistic versus decomposed PAQ job dimension ratings approached zero, even for raters who were highly skilled in the use of the PAQ. A similar study by Harvey, Wilson, and Blunt (1994) reported generally quite low levels of convergence between holistic versus decomposed JA ratings in a task-inventory analysis.

In the domain of job specifications, relatively little research has focused on directly comparing the convergence of holistic ratings of AO traits against independently validated standards, however, the results of DeNisi and Shaw (1977) are quite consistent with those seen in the JA domain, and should represent serious cause for concern for anyone advocating widespread use of holistic ratings of AO traits. In this study, holistic self ratings of ability-trait constructs were compared against scores on the same constructs that were obtained via traditional (decomposed) multi-item assessment tests; paralleling the Butler and Harvey (1988) results, DeNisi and Shaw reported rates of convergence that approached zero.

7. There is no reason to believe that a single “true” profile of hypothetical AO trait requirements exists for any job or occupation. One of the most troublesome assumptions that underlies methods of “job analysis” – or, more

correctly, methods of inferring job specifications – that purport to be able to holistically rate the AO-based worker trait requirements of jobs or occupations (e.g., Fleishman, 1992; Fleishman & Mumford, 1991; Fleishman & Reilly, 1992; Fleishman, Wetrogan, Uhlman, & Marshall-Mies, 1995; Hughes & Prien, 1989; Lopez et al., 1981) is the notion that a single “correct” or “optimal” profile of ability traits exists for each job (an equally troublesome assumption – namely, that such an ideal profile can be identified using single-item holistic rating scales – was discussed in the previous section). The definition of JA advanced by Sanchez and Levine clearly includes holistic rating of AO-based worker trait requirements as a “job analysis” method, and it is effectively essential that one must make such an assumption when using such techniques, especially when one views the group mean rating as representing the best approximation of the “true” profile of the job in question; a similar assumption is likewise required when SMEs are asked to holistically identify the optimal profile of worker traits when conducting “competency modeling” (e.g., Sanchez & Levine, p. 10).

The fundamental flaw that we find with this view is that it assumes that it is effectively impossible to find a situation in which compensatory relations exist between the worker-attribute requirements identified in the JS process. For example, in a job requiring both Mechanical Reasoning and Trunk Strength, it is easy to imagine a situation in which two employees exhibit identical overall job effectiveness, yet they have different profiles of scores on these two traits (i.e., higher strength compensating for lower cognitive skills, and vice versa). As the number of required KS and AO traits increases, the number of possible different “successful profiles” across these traits increases geometrically. If compensatory relations are possible, this causes obvious – and fundamental – problems for holistic strategies of inferring AO profiles.

8. “Consequential validity” sounds like “utility.” Although we’re not irrevocably opposed to suggestions that we rename job analysis to something more appealing (e.g., Levine & Sanchez, 1999, suggested renaming it “work analysis”) as a means to address its image problem (Cunningham, 1989), we cannot accept

Sanchez and Levine's conclusion that the only truly useful way to "evaluate job analysis accuracy... [is by] focusing on the consequences of job analysis data" (p. 2), or the "consequential validity" (p. 1) of applying JA data to solve specific organizational problems.

We identify two main issues here. First, although we agree that validity is fundamentally an issue of correctness of inference, the so-called "consequential validity" discussed by Sanchez and Levine seems to bear a much closer correspondence to classical notions of utility than a synonym for job analysis accuracy. We see no point in coining a new term to describe an already widely used concept. Second, the direction of causality seems to have been reversed: at several points, they essentially argue that job analysis accuracy is caused by the way in which such data are used or applied to solve organizational personnel problems. For example, "accuracy does not exist in a vacuum but it is defined precisely by the consequences of [using] such data" (p. 16, emphasis added), and "we frame the notion of accuracy in JA in a different light: one that focuses on its consequences" (p. 4, emphasis added). From such an analysis, it would seem as though any number of counterintuitive situations could occur: for example, (a) if two identical organizations collected identical JA databases, the same data could be deemed "inaccurate" or "invalid" if used by one organization for a given purpose (e.g., identifying an ability-oriented selection test) yet "accurate" or "valid" for the other in a different context; or (b) in a different organization, a completely inaccurate and erroneous JA database (e.g., collected anonymously by certifiably compulsive liars) could be found to have "consequential validity" if by pure chance the inferences drawn from it (e.g., that a given ability test would produce a statistically significant r when predicting job performance) were found to be useful in practice.

We would contend that the opposite direction of causality is more realistic: namely, that the probability of finding consequential validity or utility when applying JA data to make a specific personnel inference (e.g., which selection test to use) is going to be the result of the intrinsic quality and accuracy of the JA database, and not the reverse. In essence, as we described in

Figure 3, the quality and accuracy of the JA ratings is a property that is in no way dependent on the way in which practitioners may subsequently attempt to apply it to make personnel decisions and inferences. If the JA database is accurate, the results of applying it may or may not be found to be useful and valid in practice, but that will be determined entirely by the quality and correctness of the inference process that is followed when the practitioner attempts to link the JA ratings to the desired personnel decision. Clearly, these decisions are not properties of the JA database, but rather depend entirely on the skill, expertise, and strategy employed by the personnel practitioner.

9. "True score" JA research studies can indeed be useful. We disagree with Sanchez and Levine's assertions (e.g., pp. 5-11) that JA researchers must abandon the practice of conducting studies that evaluate JA accuracy by focusing on the degree of discrepancy between a known "standard" or "true score" and the JA ratings provide by some group of interest. Clearly, the designs of some studies of this ilk were indeed so highly contrived – for example, using highly unrealistic JA inventories, job-naïve raters, or unrealistically brief and atypical JA rating modalities such as making ratings after watching a few minutes of video taped job performance – that serious questions could be raised regarding their external validity and generalizability to the "real world" (e.g., Arvey, Davis, McGowen, & Dipboye, 1982; Arvey, Passino, & Lounsbury, 1977; Jones, Main, Butler, & Johnson, 1982; Stone & Gueutal, 1985; Wexley & Silverman, 1978). However, this fact does not mean that useful data has not been produced by other studies performed in more realistic conditions, or that this paradigm cannot be useful in the future.

Of course, we must note that studies of the JA (or JS) rating process that define inaccuracy as a deviation from a known-true criterion do indeed suffer from a different type of limitation: namely, that they are primarily valuable only when they show that a given rating methodology does not perform well. That is, when a given type of JA or JS rating method has been convincingly shown to exhibit unacceptable levels of convergence with the known criterion of accuracy in realistic rating

situations, practitioners would be wise to conclude that the use of such a method would be ill-advised. For example, there was a period of time in which some authors (e.g., Jones et al., 1982) claimed that practitioners could collect detailed worker-oriented JA ratings using very little effort; in this case, by using job-naïve raters who had been given only a very brief narrative summary of the job's duties. It took a series of "true score" studies and studies using Monte Carlo methods (e.g., Butler & Harvey, 1988; Harvey & Lozada-Larsen, 1988; Harvey & Hayes, 1986) to empirically demonstrate what many thought should have been intuitively obvious in the first place: namely, that job-naïve raters cannot be relied upon to provide accurate JA ratings, and that their inability to converge with a known criterion generalizes across a wide range of rating situations.

Unfortunately, the converse does not necessarily hold: i.e., even if hundreds of tightly controlled lab studies using unquestionably correct "true scores" as their criterion were able to show that a given JA rating instrument or technique produces high convergence with a known accuracy criterion, practitioners could not reasonably or necessarily assume that equally good levels of performance would be found when different raters, rating different (or even the same) jobs, would use the JA method in a new situation. As we discuss below, the issue of JA accuracy must always be settled on a situation-by-situation basis; even if a given rating method has been successfully used in thousands of previous real-world settings, this fact proves nothing regarding the quality and accuracy of JA ratings (or the validity of JS speculations) obtained using that method in any new setting.

10. There is no criterion problem in job analysis; however, the "cognitive lab study" paradigm is not the answer to JA accuracy and JS validity questions. Not surprisingly, we also disagree with Sanchez and Levine's claim that past JA research has been of such limited value because "the absence of sound criteria plagues the JA domain" (p. 9). Indeed, we would advance precisely the opposite conclusion: namely, that in JA, there is no criterion problem (although there certainly is one in the domain of making speculative JS ratings of ability trait

requirements; e.g., due to the likelihood of finding compensatory relations when multiple worker AO-traits are used as predictors of job performance). As was discussed above, when sound, verifiable Type III JA rating methods are used, it is a very straightforward matter to identify a certifiably accurate profile of JA ratings for any given position via independent SME review and verification.

The "catch" is, of course, that this (a) requires significant effort (e.g., it may take more time to review ratings and fix mistakes than it took to collect the first round of ratings); and (b) JA accuracy must be demonstrated in each situation in which the JA is performed. As we discussed earlier, it has consistently been our experience that when one takes the trouble to actually conduct an independent SME review and verification of JA ratings, one invariably finds widespread – and often serious – rating inaccuracies. Such errors would never be caught in the absence of an independent review, and the fact that we have observed this phenomenon to occur consistently across a very wide range of JA situations leads us to the conclusion that the accuracy of JA ratings can never be taken for granted in any new situation in the absence of an independent accuracy review.

We therefore see little point in the current enthusiasm for "cognitive analysis" of the job rating process. For real-world JA, what matters is identifying and resolving rating inaccuracies; unfortunately, all the carefully controlled, cognitive theory-driven, laboratory studies in the world – unless they show that a given rating methodology produces consistently unacceptable data – are of virtually no practical importance when one's task is to defend the accuracy of the JA ratings of this job, in this situation, or to document the validity of the JS inferences made in this situation on the basis of those JA ratings. In essence, the cognitive, social, political, or motivational causes that may contribute to observed JA inaccuracies matter little in practice; what matters is finding and fixing those inaccuracies, whatever their causes may have been.

At times we get the impression when reading articles that advocate an ambitious program of cognitive-theory-based research on the JA and JS

rating process (e.g., Morgeson & Campion, 1997) that at the end of this process, if everything works well, it is expected that researchers will be able to identify a set of “magic bullet” JA and JS rating methodologies that we can count on to produce universally accurate, valid results (given their critical comments regarding lab studies of JA, this may also represent an area of agreement with Sanchez & Levine, 1999). In our assessment, this is a completely misplaced hope: i.e., despite the fact that in JA there can indeed be an objective reality and “gold standard” of accurate JA ratings for each position (if you’re willing to expend the effort required to identify it, that is), the accuracy of JA ratings must always be demonstrated on a case-by-case, position-by-position basis. Barring the unlikely possibility that organizations will someday acquire the ability to exert absolute control over all aspects of what every worker on a given job does, we see no chance that there will ever be a form of “accuracy generalization” – analogous to validity generalization (VG) – in which practitioners can simply take the results of past job analyses, or use JA methods that were previously found to work well, and magically generalize the accuracy found for a given job (or JA rating method) in a previous situation to a new situation of interest.

In essence, job analysis represents the last bastion of true situational specificity (SS). Although it may or may not eventually be demonstrated convincingly that cognitive ability test validities are significantly moderated due to SS (e.g., Schmidt et al., 1981), there is every reason to believe that large-scale – and true – differences exist across the different situations (e.g., geographic location, industry type, unionized versus nonunion) in which a given job or occupational title may exist. As was noted earlier, even in a single organization it is common to find very significant cross-position variability within a given job title (e.g., an Administrative Assistant in the Public Works department might have significantly different duties than an Administrative Assistant in the Police Department). The same phenomenon, although probably on a much larger scale, also occurs when the supposedly same job or occupational title is described across different organizations, geographic areas, or industries.

Indeed, the effort to dramatically reduce the number of occupational titles recognized by the US Department of Labor as part of their project to replace the DOT with the online O*Net database of occupational information (e.g., Peterson et al., 1997) will tend to exacerbate the problem of cross-situational, within-title variability. That is, the byproduct of collapsing approximately 13,000 previous occupational titles into only a couple of thousand new titles will almost unavoidably be that the new occupational titles will have much higher levels of true within-title, cross-situational variability in work activities. Very little research attention has been paid to the issue of both cross- and within-setting heterogeneity; particularly with respect to determining the degree to which this true cross-position variance moderates the validity of decisions based on those JA ratings (especially, in terms of AO-based job specifications or “competency” requirements). In our assessment, this topic – and not the search for “magic bullet” methods of JA or JS rating that will demonstrate “accuracy generalization” – should be our top research priority.

11. Final issues. Although the reader may be tempted to conclude by this point that JA researchers and practitioners already have plenty to worry about, we close by noting that there are additional, equally troublesome issues of accuracy and validity in the JA rating process that did not even indirectly make it onto the list of concerns raised by Sanchez and Levine. We will briefly touch on three that we find to be important.

First, the test-retest stability of JA ratings (typically, ratings of behaviorally detailed task statements) made by incumbents has not been found to be especially good, even under the best of circumstances (e.g., when incumbents re-rate a number of tasks within the same inventory to provide a type of repeat-item “reliability”). Although these “test-retest reliabilities” vary somewhat in magnitude depending on how they are calculated (e.g., Wilson, Harvey, & Macy, 1991; Murphy & Wilson, 1998), they are typically not of the magnitude that one would view as being adequate using generally accepted standards of reliability (as was noted above, it is not uncommon to find that a task receives a DNA rating on one occasion and is rated as being

highly applicable on the repeat rating). Rather than providing additional evidence against the existence of an objective reality, we view such findings as indicating that even under extremely favorable conditions (i.e., a single-sitting administration of an easily understandable and verifiable JA instrument), incumbents – without the certainty of independent SME review of the ratings – cannot be counted upon to provide demonstrably accurate JA ratings. Given this disturbing level of within-situation variance observed in task ratings, one hesitates to speculate regarding the degree to which raters would produce inconsistent ratings when asked to rate more abstract, less-verifiable aspects of work (or AO-based worker-trait requirements) when there is similarly no possibility for “true” changes in the characteristics being rated over the duration of the rating process. The only study of which we are aware that has addressed this question showed that test-retest stability of incumbent ratings declines in inverse proportion to the degree of behavioral abstraction in the items being rated (Murphy & Wilson, 1998), raising further questions regarding the advisability of using more holistic rating tasks. In our assessment, these are the types of “rating process” issues that are in need of much additional research.

Second, the issue of attempting to quantify the validity (in a content-coverage sense) of a job analysis inventory based on subjective incumbent perceptions is equally troublesome. In short, incumbents seem to be incapable of estimating the proportion of the entire set of tasks they perform that is present in a given list of tasks they are asked to rate. In the only study of which we are aware that addressed this issue, incumbents on average indicated that 75% of the tasks they performed were still present in a deliberately reduced inventory, even when over two-thirds of the tasks that were known to be applicable based on the results of a previously rated inventory had been removed from the second survey (Wilson, 1996). What is particularly disturbing about these results is that they are based on experienced job incumbents, rating easily verifiable, job-specific tasks. In short, if such raters cannot reliably rate tasks or estimate the percentage of a job’s content domain that is covered by a given task-inventory, how can we expect other SMEs –

especially, those who have not been exposed to the complete list of tasks, as the above raters were – to do better when rating the degree of content-coverage of more abstract JA or JS items?

Finally, it is significant to note the absence of any discussion in Sanchez and Levine of empirically based, non-judgmental methods for linking JA ratings to JS inferences. The methods they discussed for identifying job specifications and other end-uses of JA data primarily involved either direct inferences of the KS- or AO based constructs to be used for the job specifications, or an apparently rational process in which “the organization adopts specific rules concerning the making of linkages between JA data and JA-based organizational products” (pp. 14-15, and their Figure 1). In our assessment, one of the most promising methods (and best-kept secrets) in industrial psychology is what McCormick termed job-component validity, or JCV (e.g., Brown & Harvey, 1996; Cunningham, 1964; McCormick et al., 1972), which involves statistically capturing the ways in which KS- and AO based worker-traits are predictable from the quantitative scores of their jobs on an array of moderate- or low-specificity job dimensions (the JA items in the top portion of Figure 1). Once these predictive relations are known, the worker-attribute requirements of new jobs can be empirically forecast by applying the prediction equations to the JA results for the new jobs in question. Although we certainly do not maintain that JCV represents the “magic bullet” that will solve all of our worker-trait specification problems, in our assessment the use of an empirical method based firmly in the results of a defensible JA is vastly preferable to relying on subjective “linking rules” to guide the process.

Summary Conclusions

In sum, we agree fully with Sanchez and Levine regarding the critical importance of following formalized procedures for linking JA data to JS inferences, collecting empirical evidence to justify the “consequential validity” of worker-trait inferences, and the limitations of highly contrived “lab” studies of the JA and JS processes. However, many areas of fundamental disagreement remain, and in our assessment they represent far more than esoteric academic quibbling over terminology; in many respects, our

visions of the job analysis process are diametrically opposed. Their approach (a) intermixes ratings of work activities with inferences of hypothetical worker-trait requirements under the common label of “job analysis;” (b) argues against attempts to directly assess the accuracy of JA data under the assumption that it “cannot be determined in the absence of knowledge of [consequential validity] effects” (p. 10); and (c) rejects the notion that a “gold standard” of JA accuracy exists in favor of a relativist ontological view that tolerates even very large levels of within-title disagreement as reflections of “alternate but equally valid views” (pp. 7-8) of the job. In contrast, we (a) focus on rating specific, verifiable aspects of work behavior and job context in the JA process (i.e., the Type III methods in Figure 2, consisting of well-constructed task- and worker-oriented instruments); (b) believe that such JA ratings can – and must – be subjected to independent SME review to verify their accuracy and resolve rating errors; and (c) recommend constructing job-title systems that minimize true within-title heterogeneity and problems due to aggregation bias prior to using such data to drive personnel functions.

We are deeply troubled by the implications of accepting a relativist view that holds that reality is socially constructed, and that JA accuracy can never be directly assessed. How does one make scientifically based prescriptions when one freely acknowledges that there may be an infinite number of “equally valid” – but contradictory – realities, and asserts that there is no way to directly determine whether a given position’s ratings are accurate or not? As a practical matter, what are the legal-defensibility implications of adopting the view that practitioners need not concern themselves with directly assessing the accuracy of the JA database that forms the foundation for an array of highly litigated personnel decisions (e.g., selection, promotion, placement, compensation), or its corollary, that the presence of high levels of cross-rater, within-job disagreement does not necessarily represent cause for concern?

We contend that our “bounded positivism” represents a much more sound position from which to undertake a scientific consideration – or

a courtroom defense – of the JA and JS process than a relativistic approach that denies that an objective reality in the world of work even exists. Instead of gambling that subsequent uses of the JA database to drive personnel decisions will provide indirect evidence regarding the accuracy of the JA ratings, we suggest that personnel practitioners should focus first and foremost on demonstrating the accuracy of their JA ratings (including the issue of the domain-sampling adequacy of the JA instruments). If this can be demonstrated, one can then proceed to document the process that leads from the JA database to each personnel function derived from it (ideally, by relying on empirically based linking methods such as JCV when setting employee selection requirements), and assess the bottom-line utility (or “consequential validity”) of these applications.

However, if the accuracy of the JA database cannot first be documented and verified, we see little point in continuing the process. In our view, both the legal defensibility as well as the bottom-line effectiveness of personnel applications derived from JA data will unavoidably be a joint function of three major factors: (a) the accuracy of the JA database, (b) selecting the correct level-of-specificity and type of JA data to drive each personnel function (see Figure 1), and (c) following an appropriate set of decision rules and empirical procedures to link the JA data to each subsequent personnel function. Failure at any of these three steps would likely doom the entire process to failure. As such, we flatly reject the implication that one is left with by Sanchez and Levine’s arguments that somehow a bottom-line defense of “consequential validity” (e.g., showing a statistically nonzero relation between a selection test and job performance) will be capable of compensating for the presence of gross cross-rater disagreement or inaccuracy in the JA data, or that an effective personnel system (e.g., selection, training) can be derived – other than by pure chance – through the application of fundamentally inaccurate JA ratings.

References

American Educational Research Association, American Psychological Association, & National

- Council on Measurement in Education (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.
- Arvey, R. D., Davis, G. A., McGowen, S. L., & Dipboye, R. L. (1982). Potential sources of bias on job analytic processes. Academy of Management Journal, 25, 618-629.
- Arvey, R. D., Passino, E. M., & Lounsbury, J. W. (1977). Job analysis results as influenced by sex of incumbent and sex of analyst. Journal of Applied Psychology, 62, 411-416.
- Brown, R. D., & Harvey, R. J. (1996, April). Job-component validation using the MBTI and the Common-Metric Questionnaire (CMQ). Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, San Diego.
- Butler, S. K., & Harvey, R. J. (1988). A comparison of holistic versus decomposed rating of Position Analysis Questionnaire work dimensions. Personnel Psychology, 41, 761-771.
- Christal, R. E. (1974). The United States Air Force occupational research project (AFHRL-TR-73-75). Lackland AFB, TX: Air Force Human Resources Laboratory, Occupational Research Division.
- Cornelius, E. T., & Lyness, K. S. (1980). A comparison of holistic and decomposed judgment strategies in job analyses by job incumbents. Journal of Applied Psychology, 65, 155-163.
- Cragun, J. R., & McCormick, E. J. (1967). Job inventory information: Task reliabilities and scale interrelationships (PRL-TR-67-15). Lackland AFB, TX: Personnel Research Laboratory. (NTIS No. AD-681-509).
- Cunningham, J. W. (1964). Worker-oriented job variables: Their factor structure and use in determining job requirements. Unpublished doctoral dissertation, Purdue University.
- Cunningham, J. W. (1989, August). Discussion. In R. J. Harvey (Chair), Applied measurement issues in job analysis. Symposium presented at the American Psychological Association convention, New Orleans.
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. Journal of Applied Psychology, 78, 98-104.
- Crocker & Algina (1986). Classical and modern test theory. Orlando: Harcourt Brace Jovanovich.
- Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. Psychological Bulletin, 81, 95-106.
- DeNisi, A., & Shaw, B. (1977). Journal of Applied Psychology.
- Fleishman, E. A. (1992). Rating scale booklet: F-JAS: Fleishman job analysis survey. Palo Alto, CA: Consulting Psychologists Press.
- Fleishman, E. A., & Mumford, M. (1991). Evaluating classifications of job behavior: A construct validation of the ability requirement scales. Personnel Psychology, 44, 253-575.
- Fleishman, E. A., & Reilly, M. E. (1992). Administrator's guide: F-JAS: Fleishman job analysis survey. Palo Alto, CA: Consulting Psychologists Press.
- Fleishman, E. A., Wetrogan, L. I., Uhlman, C. E., & Marshall-Mies, J. C. (1995). Abilities. In Development of prototype occupational analysis network (O*NET) content model: Volume I: Report. Utah Department of Employment Security.
- Gutman, A. (1993) EEO law and personnel practices. Sage.
- Harvey, R. J. (1990, April). Incumbent versus supervisor perceptions of job tasks. In K. Kraiger (Chair), Cognitive representations of work. Symposium presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Miami.
- Harvey, R. J. (1991). Job analysis. In M. D. Dunnette and L. M. Hough (Eds.), Handbook of Industrial and Organizational Psychology (2nd ed., pp. 71-163). Palo Alto, CA: Consulting Psychologists Press.
- Harvey, R. J., & Hayes, T. L. (1986). Monte carlo baselines for interrater reliability correlations using the Position Analysis Questionnaire. Personnel Psychology, 39, 345-357.
- Harvey, R. J., & Lozada-Larsen, S. R. (1988). Influence of amount of job descriptive information on job analysis rating accuracy. Journal of Applied Psychology, 73, 457-461.

Harvey, R. J., Wilson, M. A., & Blunt, J. H. (1994, April). A comparison of rational/holistic versus empirical/decomposed methods of identifying and rating general work behaviors. Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Nashville.

Hughes, G. L., & Prien, E. P. (1989). Evaluation of task and job skill linkage judgments used to develop test specifications. Personnel Psychology, *42*, 283-292.

James, L. R. (1982). Aggregation bias in estimates of perceptual agreement. Journal of Applied Psychology, *67*, 219-229.

Jones, A. P. Main, D. S., Butler, M. C., & Johnson, L. A. (1982). Narrative job descriptions as potential sources of job analysis ratings. Personnel Psychology, *35*, 813-828.

Lopez, F. M., Kesselman, G. A., & Lopez, F. E. (1981). An empirical test of a trait-oriented job analysis technique. Personnel Psychology, *34*, 479-502.

McCormick, E. J., DeNisi, A., & Shaw, B. (1979). Use of the Position Analysis Questionnaire for establishing the job component validity of tests. Journal of Applied Psychology, *64*, 51-56.

McCormick, E. J., Jeanneret, P. R., & Mecham, R. C. (1972). A study of job characteristics & job dimensions as based on the Position Analysis Questionnaire (PAQ). Journal of Applied Psychology, *56*, 347-368.

Pass, J. J., & Robertson, D. W. (1980). Methods to evaluate scales and sample size for stable task inventory information. Navy Personnel Research and Development Center: NPRDC TR 80-28.

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., Fleishman, E. A., & Levin, K. Y. (1997). O*Net final technical

report. Utah Department of Workforce Services, Contract Number 94-542.

Sackett, P. R., Cornelius, E. T., & Carron, E. T. (1981). A comparison of global judgment vs. task-oriented approaches to job classification. Personnel Psychology, *34*, 791-804.

Smith, J., & Hakel, M. D. (1979). Convergence among data sources, response bias, and reliability and validity of a structured job analysis questionnaire. Personnel Psychology, *32*, 677-692.

Society for Industrial and Organizational Psychology, Inc. (1987). Principles for the validation and use of personnel selection procedures (Third Edition). College Park, MD: Author.

Stone, E. F., & Gueutal, H. G. (1985). An empirical derivation of the dimensions along which characteristics of jobs are perceived. Academy of Management Journal, *28*, 376-396.

Stutzman, T. M. (1983). Within classification job differences. Personnel Psychology, *36*, 503-516.

Uniform guidelines on employee selection procedures (1978). Federal Register, *43*, 38290-38315.

Webb, N. M., Shavelson, R. J., Shea, J., & Morello, E. (1981). Generalizability of general education development ratings of jobs in the United States. Journal of Applied Psychology, *66*, 186-192.

Wilson, M. A., & Harvey, R. J. (1990). The role of relative-time-spent ratings in task-oriented job analysis. Journal of Business and Psychology, *4*, 453-461.

Wilson, M. A., Harvey, R. J., & Macy, B. (1990). Repeating items to estimate the test-retest reliability of task inventory ratings. Journal of Applied Psychology, *75*, 158-163.

Table 1
Possible Methods for Forming an Aggregate Job Profile

Number of Positions Rated	Number of Raters of Each Position	
	one	multiple
one	<ul style="list-style-type: none"> - Type A - “Gold standard” of accuracy can be achieved by item-level review and editing of ratings by qualified, objective SME (“validation”) <u>if</u> items and rating scales are sufficiently verifiable - Position ratings that have not been validated cannot be presumed accurate 	<ul style="list-style-type: none"> - Type B - Cross-rater, within-position agreement directly indexes accuracy, to the extent that raters are qualified, objective SMEs - All disagreement should be removed by further review prior to forming job-aggregate profile because no “true” cross-rater within-position disagreement is possible - All cross-rater, within-position disagreement must reflect (a) lack of adequate job knowledge in raters, or (b) flaws in the items or rating scale (imprecise, holistic)
multiple	<ul style="list-style-type: none"> - Type C - Cross-rater agreement does <u>not</u> imply “reliability” or accuracy if ratings were not made or validated by qualified, objective SMEs - If position ratings are validated, cross-position <u>agreement</u> implies accuracy of the aggregate description - If validated, cross-position <u>disagreement</u> implies job-title-system problems (i.e., presence of true, significant cross-position activity differences) - If position ratings are not validated, cross-position agreement does not imply accuracy, nor does cross-position disagreement imply lack of accuracy - Aggregate profile will be subject to <u>aggregation bias</u> when nontrivial cross-position differences exist, regardless of reason 	<ul style="list-style-type: none"> - Type D - Same issues as Type II regarding within-position disagreement, and as Type IV regarding cross-position agreement

Figure 1. Illustration of levels-of-analysis issues in linking the domain of human skill/ability constructs with the domain of work activities; illustrative items are listed from most specific (bottom) to most abstract (top), with the dashed vertical line in the middle representing the “inference barrier” that must be bridged when linking job analysis data to AO-based worker-trait requirements or KS-based skill requirements.

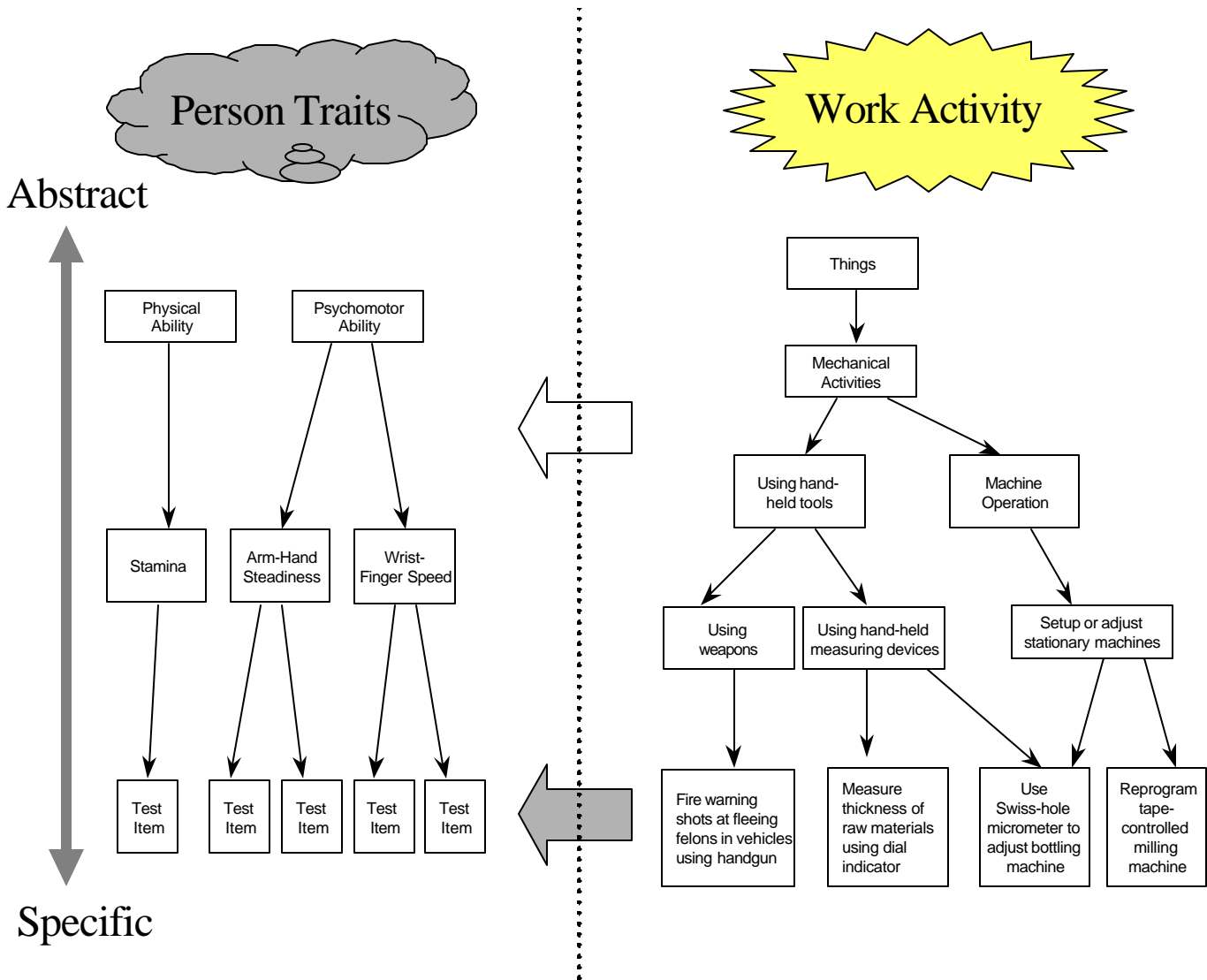


Figure 2. Possible types of job analysis ratings, based on the combination of rating scale metric versus level of behavioral/technical specificity of the work activity item. Because the presence of either a relativistic rating scale or a holistic work activity item (but especially, the presence of both) raise serious questions regarding the verifiability of the resultant rating, we argue that Type IV ratings should be used to collect job analysis data.

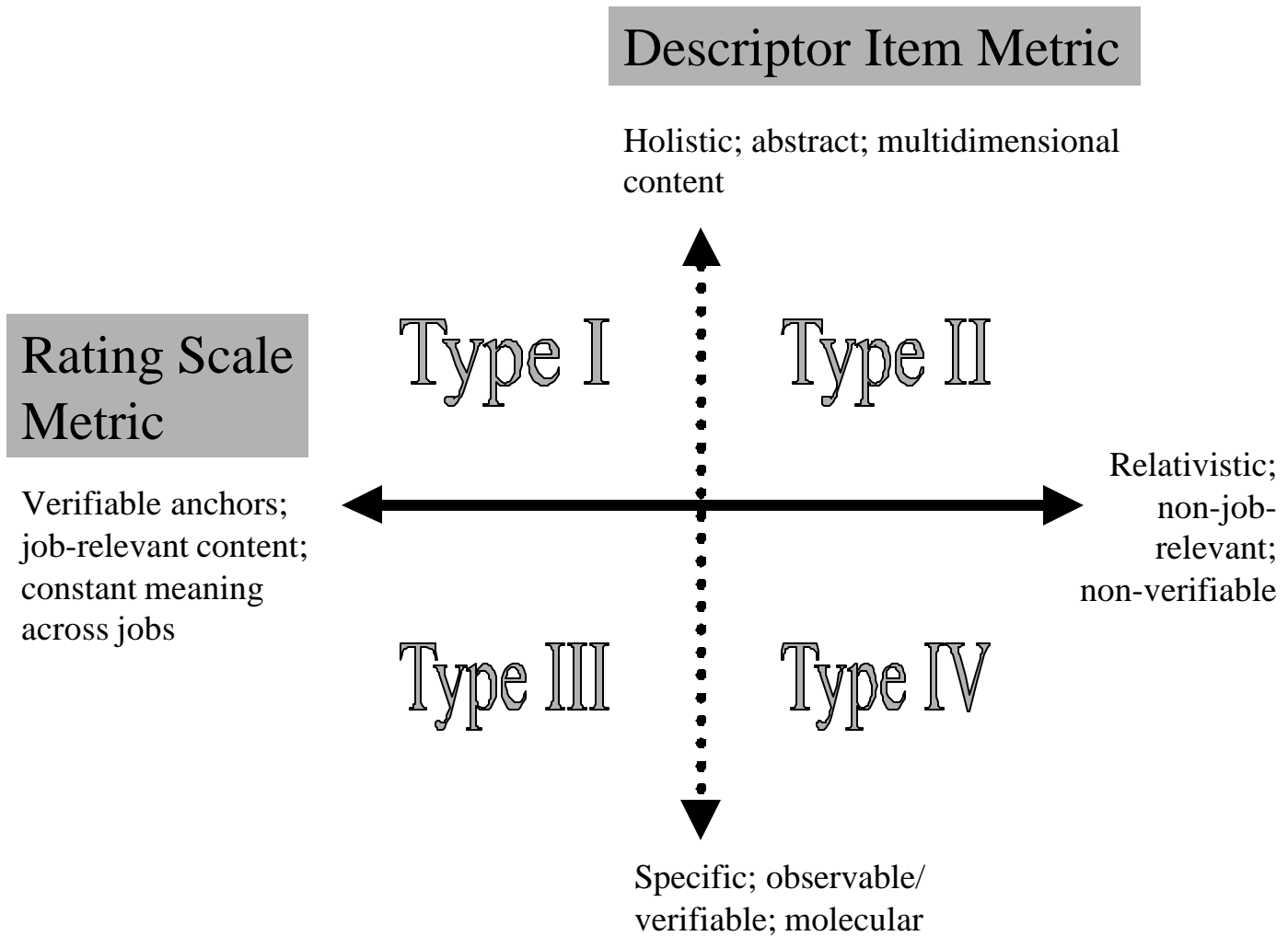
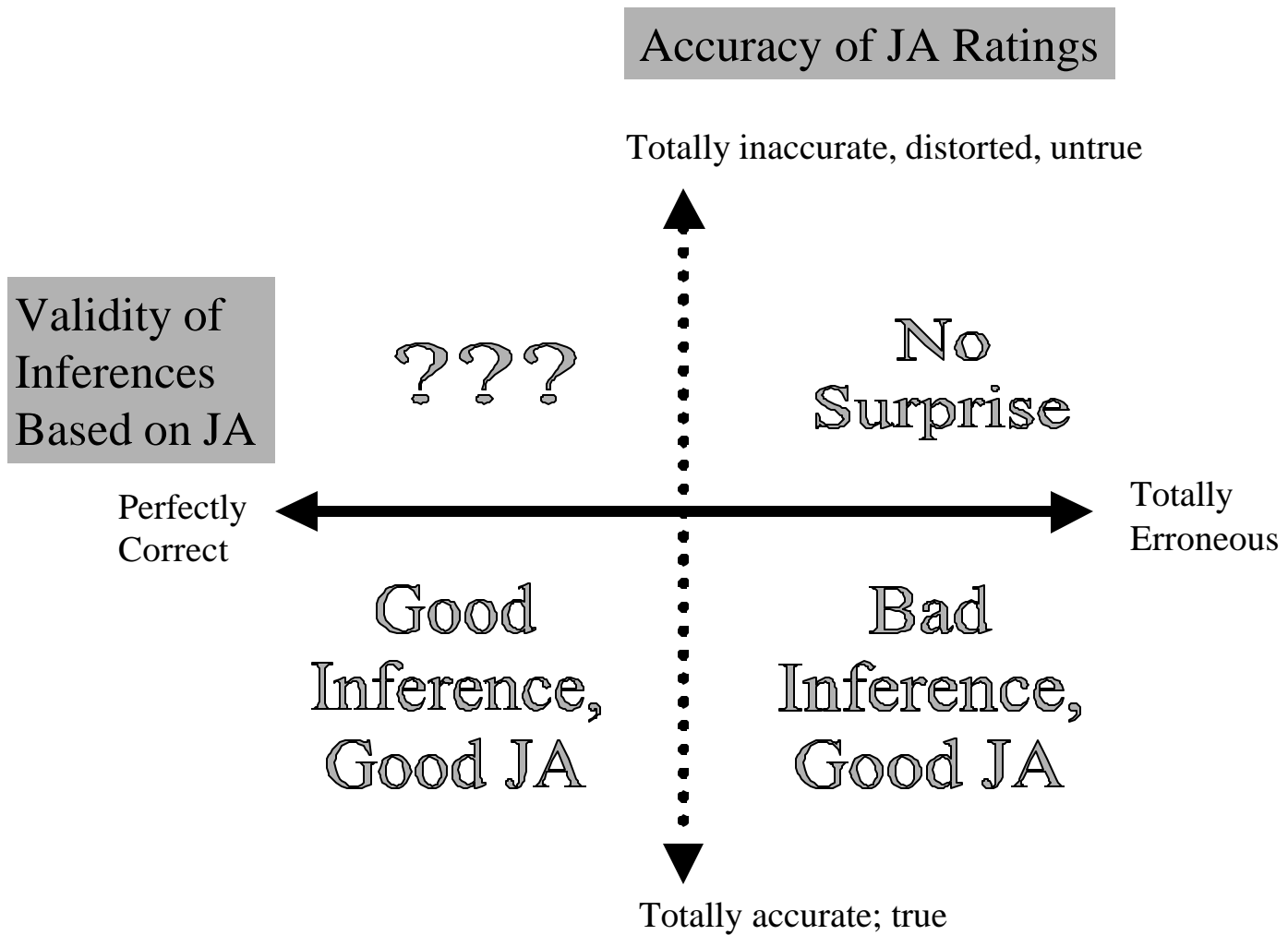


Figure 3. Possible outcomes when making personnel inferences (e.g., KS, AO requirements) from job analysis ratings, varying as a function of the quality of the job analysis ratings and the degree to which the inferential leap from the job analysis data to the personnel requirement is valid.



Author Note. Address correspondence to: R. J. Harvey, Department of Psychology, Virginia Tech, Blacksburg, VA 24061-0436. Phone (540) 231-7030, email: harveyrj@vt.edu.