PART I

Foundations

CHAPTER ONE

History of Research Methods in Industrial and Organizational Psychology: Measurement, Design, Analysis

James T. Austin, Charles A. Scherbaum, and Robert A. Mahlman

Our aim in this chapter is to review the history of research methods. An underlying premise is that in so doing we can improve current research. Research methods and theories enable the description, prediction, and understanding of organizational behavior. Phenomena of longstanding concern to industrial and organizational (I-O) psychologists (Wilpert, 1997) pertain to broadly-defined behavior by groups and individuals, within organizations, and the interrelationships among these levels. It is clear that the evolution of research methods brought the wealth of choices available to I-O psychologists (MacCallum, 1998; Sackett and Larson, 1990). What important issues arise from studying the history of research methods? We used three strategies to accomplish this review. We examined published research and historical discussions (e.g., Cowles, 1989; Morawski, 1988; Owen, 1976); in addition, we coded 609 empirical articles over a period of time from the *Journal of Applied Psychology* (*JAP*) in order to track researcher choices (cf. Sackett and Larson, 1990).

A Time X Domain framework organizes this chapter. The levels of the time facet are intervals that span 1904 to 1935, 1936 to 1968, and 1969 to 2000. Intervals were selected to group time and also to identify landmarks and trends. In the year 1935, for example, Thurstone and colleagues founded the Psychometric Society and Fisher published *Design of Experiments*. In 1968, *Statistical Theories of Mental Test Scores* (Lord and Novick, 1968) appeared and Cohen (1968) brought the general linear model into wider view. Currently, there are several potential landmarks in the research methods literature. One is computational modeling (Ilgen and Hulin, 2000), another is an integration of test theory models (McDonald, 1999), and a third consists of volumes honoring and extending the work of Donald Campbell (Baum and McKelvey, 1999; Bickman, 2000a, 2000b) and Douglas Jackson (Goffin and Helmes, 2000).

Table 1.1 Representative important developments within a methods domain X temporal interval matrix

Method domain	*Temporal interval facet*		
	1904–1935	*1936–1968*	*1969–2000*
Measurement	Classical test theory	Item response theory, construct validity	Generalizability theory, consequential validity
Design	Longitudinal, design of experiments	Internal, external validity	Multi-level designs, causal inference
Analysis	Inference, multiple regression analysis	Multivariate methods ANOVA/ANCOVA	Quantitative synthesis, covariance structure models

The levels of the second facet, research methods, are measurement, design, and analysis (Pedhazur and Schmelkin, 1991). *Measurement* consists of conceptualizing and scoring the attributes of entities. *Design* involves planning and executing research to support valid inferences that may generalize beyond the sample. *Analysis* is making sense of the resultant data from measurement and design. Choices made in all these domains clearly influence study interpretations. Table 1.1 presents the organizing framework with one or more representative developments for each domain-interval intersection.

Although presented orthogonally, the levels of the research methods facet are indeed closely interrelated. Consider warnings against substantive research without first establishing construct validity (Schwab, 1980). Consider situations in which poorly designed research aims to inform policy. Lastly, consider the clash about whether analysis depends upon level of measurement (Michell, 1986, 1999). We force the separation of levels for purposes of exposition.

Historically, research methods first developed in wider spheres. We thus provide a general history of each domain using the time intervals as rough boundaries before tracing developments within the I-O field. In part, this tactic recognizes that, during the formative years of I-O psychology, researchers and practitioners were trained in experimental psychology (Katzell and Austin, 1992). In fact, Walter Dill Scott and Hugo Münsterberg were trained in Europe by Wilhelm Wundt, a founder of experimental psychology. In part, this tactic recognizes the role of the diffusion of innovations, a process by which innovations spread through various information channels over time (Johns, 1993; Rogers, 1995). The process of innovation helped research methods permeate into new areas of psychology, in this case from experimental to industrial psychology. In general, innovations in measurement, design, and analysis have diffused – sometimes slowly, sometimes more quickly – from developers of research methods to researchers who actually implement those methods. We first describe our coding of *JAP* articles. Although not the only journal of the field, it does have a long publication history in the United States, and this figured in our choice. Our purpose is to amplify discussion in the sections that follow on measurement, design, and analysis.

Snapshots over Time from the *Journal of Applied Psychology*

Consider the first empirical article in *JAP*. Terman (1917) evaluated 30 candidates for municipal positions using a shortened Stanford-Binet (S-B) and 7 "pedagogical" tests. He described his sample (age, education), provided frequencies for mental age, IQ, and test scores, and presented a matrix of correlations among the predictors and reported salary (range .17 to .81; probable error = .078). Spearman-Brown reliability was estimated as .69 by split-halves. Criteria were candidates' reports of past salary and occupational level. A cut-off of 80 (S-B) eliminated 10 candidates. Analyses presented correlations among predictors, and with salary and general or specific sets of the tests. Observations on individual cases concluded the report. A failure to study other assessments (medical and physical exams, moral qualities) was noted as a limitation by Terman.

To develop that snapshot further into a series, we coded 609 empirical articles from nine volumes of *JAP* (every tenth volume between 1920 and 2000). Our choice of journal was predicated on prestige and length of publication, while our sampling strategy provided representative coverage. Our coding was based on the research methods facet and on previous work (Sackett and Larson, 1990; Stone-Romero, Weaver, and Glenar, 1995). The coding sheet is reproduced in figure 1.1, and the sample and data are described in tables 1.2 and 1.3. Notice in table 1.2 the increasing proportion of what would be regarded as "current" I-O topics, the pronounced shift from single to multiple authors, and the relatively small fluctuations in median sample size. Table 1.3 provides percentage-use-indices (PUI) for each domain using the calculations of Stone-Romero et al. (1995).

Measurement

Measurement and scaling as we know them today grew from procedures used by Galton to study individual differences and by Fechner to study psychophysics. Despite its de-emphasis (Aiken, West, Sechrest, and Reno, 1990), measurement remains important in I-O education, practice, and research. The concept of measurement includes representing scientific concepts, developing instruments, and evaluating score inferences through capturing relevant and irrelevant sources of score variance. Browne (2000) used categories of mental test theory, factor analysis and related methods, and multidimensional scaling to organize his historical sketch. We discuss test theory models before turning to I-O developments.

Classical test theory (CTT) emerged from Galton's measurements of individual differences. In 1904, Spearman presented his models of "g" and measurement error. Early progress emphasized instrument development and reliability (Kuder and Richardson, 1937), with milestones being the translation-standardization of Binet and Simon's scales by Terman, origins of group testing in the alpha and beta tests (Yerkes, 1921), and creation of the army personnel system (Committee on Classification of Personnel, 1919). Books were written by Guilford (1936), Gulliksen (1950), Lord and Novick (1968), and Thurstone (1931).

Citation Information

Author(s): _____

Title: _____

Year: _____ Volume: _____ Issue: _____ Pages: _____ I-O topic?: Y N

Measurement

Measure(s): Self-report: Test Personality Attitudinal Behavioral Other _____
Behavioral: Ratings Outcomes Other ____ Physiological or other _____

Source of measures: Home-made Commercial Other _____
Reliability estimate: Y N Type: _____
Validity estimate: Y N Type: _____
Test theory: Classical IRT GT Other _____

Design

General Setting: Lab Field Simulation Other _____ Sample: _____ N _____
Design: Passive Observation Experiment Case study Archival Other _____
Temporal: Cross-sectional Longitudinal Cohort Other _____

Analysis

Descriptives used: M Md SD SE r Mo % Frequency Other _____
Primary inferential tests used: ANOVA t-tests ANCOVA MANOVA CFA EFA
Multiple Regression Path analysis Chi square MANCOVA
SEM/CSM CR PE Correlation Other _____

Statistical Conclusion Validity:
Power analysis reported N Y
Significance reported N Y
Effect size reported N Y

Figure 1.1 Protocol for coding *Journal of Applied Psychology* articles

Table 1.2 Description of the sample of *Journal of Applied Psychology* articles (N = 609)

Year	*Volume*	*Number*	*I-O topics (%)*	*Single-author articles (%)*	*Sample size statistics*		
					Mean N[a]	*Median N*	*Skew*
1920	4	30	40.0	73.3	328.20	150.0	1.079
1930	14	41	12.2	78.0	525.32	164.0	3.871
1940	24	57	24.6	77.2	3733.42	200.0	7.147
1950	34	76	60.5	51.3	580.91	118.0	4.555
1960	44	81	67.9	56.8	247.65	103.0	2.360
1970	54	85	64.7	44.7	315.94	88.00	4.560
1980	65	82	79.3	28.0	453.14	100.5	5.911
1990	75	74	94.6	27.0	2407.83	193.0	4.878
2000	85	83	85.5	08.4	1136.74	343.5	7.281

[a] Excluding meta-analysis sample sizes.

Table 1.3 Percentage of use indices (PUI) by year for measurement, design, and analysis strategies

	Measurement								
	1920	*1930*	*1940*	*1950*	*1960*	*1970*	*1980*	*1990*	*2000*
Self-Report	76.7	73.2	63.2	53.9	49.4	51.8	56.1	68.9	83.6
Behavioral	70.0	41.5	57.9	60.5	75.3	65.9	65.9	66.2	49.3
Physiological	0.0	2.4	0.0	13.2	1.2	3.5	2.4	2.7	1.5
Other	0.0	0.0	0.0	0.0	1.2	0.0	1.2	0.0	0.0
SR/Beh combination	46.7	19.5	21.1	21.1	27.2	20.0	26.8	0.0	32.8
Commercial measures	60.0	56.3	12.3	44.7	12.3	7.1	8.5	9.5	13.4
Home-made measures	36.7	39.0	56.1	44.7	64.2	70.6	62.2	37.8	38.8
Other	3.3	12.2	31.6	10.5	23.5	22.4	28.0	50.0	47.8
Classical test theory	100.0	100.0	100.0	100.0	100.0	97.6	98.8	98.6	97.0
Signal detection theory	0.0	0.0	0.0	0.0	0.0	2.4	0.0	0.0	0.0
Item response theory	0.0	0.0	0.0	0.0	0.0	0.0	1.2	1.4	3.0
Generalizability theory	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Design								
	1920	*1930*	*1940*	*1950*	*1960*	*1970*	*1980*	*1990*	*2000*
Lab	30.0	34.1	19.3	25.0	30.9	32.9	41.5	21.6	22.4
Field	70.0	63.4	80.7	71.1	67.9	63.5	53.7	63.5	65.7
Simulation	0.0	2.4	0.0	3.9	1.2	3.5	2.4	6.8	3.0
Meta-analysis	0.0	0.0	0.0	0.0	0.0	0.0	2.4	4.1	4.5
Lab/field combo	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.1	4.5
Passive-observational	70.0	78.0	61.4	60.5	51.9	45.9	43.9	52.7	53.7
Experimental	23.3	14.6	35.1	31.6	38.3	52.9	46.3	29.7	31.3
Archival	6.7	7.3	3.5	7.9	9.9	1.2	8.5	12.2	11.9
PO/exp combination	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.7	3.0
Cross-sectional	93.3	92.7	98.2	98.7	100.0	92.9	92.7	93.2	85.1
Longitudinal	6.7	7.3	1.8	1.3	0.0	7.1	7.3	5.4	14.9
	Analysis								
	1920	*1930*	*1940*	*1950*	*1960*	*1970*	*1980*	*1990*	*2000*
ANOVA	0.0	0.0	1.8	9.2	23.5	31.8	51.2	33.8	28.4
t-test	0.0	0.0	5.3	13.2	21.0	21.2	22.0	14.9	22.4
ANCOVA	0.0	0.0	0.0	0.0	0.0	1.2	2.4	4.1	4.5
MANCOVA	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.7	0.0
MANOVA	0.0	0.0	0.0	0.0	0.0	1.2	7.3	16.2	7.5
CFA	0.0	0.0	0.0	0.0	0.0	0.0	1.2	6.6	16.4
EFA	0.0	0.0	3.5	1.3	7.4	8.2	9.8	6.6	9.0
Regression	3.3	2.4	1.8	7.9	6.2	8.2	14.5	33.8	46.3
Chi-square	0.0	0.0	10.5	6.6	19.8	8.2	11.0	13.5	10.4
Correlation	0.0	0.0	0.0	14.5	24.7	23.5	35.4	37.8	25.4
Path analysis	0.0	0.0	0.0	0.0	0.0	0.0	1.2	2.7	3.0
SEM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	13.5	10.4
Critical ratio	26.7	36.6	38.6	23.7	2.5	0.0	0.0	0.0	0.0
Descriptives only	70.0	61.0	40.4	34.2	9.9	8.2	2.4	4.1	1.5
Probable error	26.7	41.5	21.1	6.6	0.0	0.0	0.0	0.0	0.0

Validity's growth spurt began during the 1950s, driven by a surfeit of terms, and continues to this day. The concept had been born as "testing the tests" (Schmitt and Landy, 1993; von Mayrhauser, 1992), first in criterion form (Scott, 1917; Toops, 1944) and later in content form (Guion, 1977). Major developments included construct validity (Cronbach and Meehl, 1955), Jane Loevinger's (1957) approach, test utility (Cronbach and Gleser, 1957), and multitrait-multimethod matrix (Campbell and Fiske, 1959). Recently Messick (1995) urged consideration of consequences as well as uses of tests through his six-part validity structure. This conception of validity is now explicit in the 1999 *Standards for Educational and Psychological Testing*. Controversies persist in the use of assessment scores to make decisions, mirroring other areas of society (Cronbach, 1975; Hanson, 1993; Herrnstein and Murray, 1994; Jensen, 1998).

Its dominance did not shield CTT from criticism (Embretson and Hershberger, 1999; Lumsden, 1976; Tryon, 1957). Two alternatives emerged during the 1950s. Item response theory (IRT) (Bock, 1997; McDonald, 1999) is a nonlinear factor model for right–wrong data rooted in Fechner's psychophysics (cf. Mosier, 1940). Lord (1952) provided the first exposition of IRT, and Lord and Novick (1968) made it accessible through their inclusion of chapters by Birnbaum. Models range from the Rasch 1-parameter to the 3-parameter, with a focus on the item and parameter invariance major advantages of IRT. Improved models address polytomous and multidimensional models (van der Linden and Hambleton, 1997). Computerized adaptive testing (CAT) uses IRT (Kyllonen, 1997; Meijer and Nering, 1999) and also provides a window on response processes. Issues for CAT-IRT, for which solutions exist, include requirements for banks of validated items, complex computer software, and the assumption of unidimensionality. Goldstein and Wood (1989) criticized IRT just as Lumsden (1976) had criticized CTT.

Cronbach and colleagues began to write a handbook and decided to study reliability first because it was well-plowed ground. Their efforts culminated in a 1972 book *The Dependability of Behavioral Measurements* (see Brennan, 1997; Cronbach, 1991). Generalizability theory (GT) liberates reliability because "error variance" is partitioned into multiple sources based on manipulating raters, items, occasions, or other facets (Shavelson, Webb, and Rowley, 1989). Inferences about the errors may be used to establish the generalizability of a construct (G-study) or to establish score usefulness in decision-making (D-study).

Several themes emerge from this review and other sources (Blinkhorn, 1997; Lewis, 1986; Traub, 1997; Wright, 1997). They include increases in (1) the complexity of models; (2) the importance of validity; (3) concern about test bias; and (4) emphasis on change measurement and predictors. What are their effects on I-O research methods?

Measurement developments in I-O psychology

Early I-O research deployed instruments using CTT. Among the instruments created were vocational interest blanks (transported from Carnegie to Stanford), Viteles' psychograph for job analysis, and the Minnesota mechanical ability tests. Thurstone (1919a, 1919b) evaluated tests predictive of telegraphy and clerical performance. Early

textbooks (Burtt, 1926; Hull, 1928; Viteles, 1932) discussed criterion validation via correlation-regression analysis. Testing was often oversold. Scaling exerted little direct influence on I-O psychology (Bass, Cascio, and O'Connor (1974) is one exception).

Between 1930 and 1945, the Great Depression and World War II provided great opportunities for I-O psychologists. During the Great Depression, job analysis research led to the *Dictionary of Occupational Titles*; Bingham published *Aptitudes and Aptitude Testing* in 1937 as a summary of available measures, and job attitudes became a research topic. By 1940, I-O psychology had come of age. Harrell's (1992) description of the Army General Classification Test and Flanagan's edited summary of Army/Air Force research are landmarks, but others helped (e.g., Stouffer and colleagues; Stuit; Cronbach and Neff). After the war, military developments were translated into business. Among them, Bray and co-workers pioneered the assessment center at AT&T, Flanagan (1954) described the critical incident technique, and Ferguson (1950) developed a large performance appraisal system. The 1960s became the era of civil rights. A practical result was equal employment opportunity and affirmative action, and a theoretical result was the emergence of test fairness and adjustment models that have continued to the present (Campbell, 1996; Cascio, Outtz, Zedeck, and Goldstein, 1991; Sackett and Wilk, 1994). It was at the end of this decade that the influence of computers for data collection and psychometric analysis increased.

I-O psychologists before 1970 were not developers, but were sophisticated users, of CTT. Since then, contributions have increased, but so too has controversy. Following its introduction, a shift toward construct validity led to James's (1973) analysis of criterion models and to Binning and Barrett's (1989) elaboration of inferential bases of validation. Other developments included presentations of psychometrics (Ghiselli, Campbell, and Zedeck, 1981) and IRT (Drasgow and Hulin, 1990), appropriateness measurement for detecting aberrant response patterns (Drasgow, 1982a), construction of measures using IRT (Drasgow, 1982b), and CAT (Drasgow and Olson-Buchanan, 1999). Sands, Waters, and McBride (1997) described the computerization of the Armed Services Vocational Aptitude Battery. Another exemplary contribution is James's use of conditional reasoning to assess personality (James, 1998). Measurement invariance, a longstanding topic, now receives increasing attention (Vandenberg and Lance, 2000). Controversies swirl around multiple topics, including general ability (Gottfredson, 1986; Sternberg and Wagner, 1993), test bias (Campbell, 1996; Hartigan and Wigdor, 1989; Gottfredson and Sharf, 1988), and testing policy issues (Wing and Gifford, 1994).

Over time, I-O psychologists have developed impressive instruments (Cook, Hepworth, Wall, and Warr, 1981). We selectively mention Functional Job Analysis (Fine, 1955), the Position Analysis Questionnaire (McCormick, Jeanneret, and Meachem, 1969), Common Metric Questionnaire (Harvey, 1993), Ability Requirement Scales (Fleishman and Mumford, 1991), and O*Net (Peterson, Mumford, Borman, Jeanneret, and Fleishman, 1999). In job attitudes, the Job Descriptive Index (Smith, Kendall, and Hulin, 1969) and the Job in General Scale (Ironson, Smith, Brannick, Gibson, and Paul, 1989) stand out, as do the Organizational Commitment Questionnaire (Mowday, Steers, and Porter, 1979) and the Affective/Continuance/Normative Commitment scales (Allen and Meyer, 1990). Well-built measures plus a counseling intervention characterize the Theory of Work Adjustment (Dawis and Lofquist, 1993). There are compilations of instruments

for organizational research (Lawler, Nadler, and Cammann, 1980; Price, 1997). In the cognitive ability domain, the Wonderlic Personnel Test has been used for screening since 1938 and Jackson's Multidimensional Aptitude Battery, which is a group-administered approach to Weschler's individually administered assessment, represents creative test construction.

Using measures requires construction as well as evaluation. Adkins and colleagues (1947) described the creation of civil service tests; Dawis (1987) and Hinkin (1995) discussed scale development; Guion (1998) presented a comprehensive discussion; and Lowman (1996) edited a sketch of the current picture. Another aspect of measurement construction concerns formats for stimulus and response. I-O psychologists have contributed since Munsterberg's use of part versus whole and Viteles' (1932) definition of analytic, analogous, and work sample assessments. Although Murray deserves credit for the assessment center, the group effort (see OSS Assessment Staff, 1948) was influenced by German psychologists (Ansbacher, 1951). The researchers at AT&T generalized the method (Howard and Bray, 1988). Computerized video testing of social-situational judgment is a current contribution (Olson-Buchanan, Drasgow, Moberg, Mead, Keenan, and Donovan, 1998). Such alternative assessments concern I-O psychologists because of possible implications for applicant reactions, adverse impact, and validity (Hakel, 1998; Reilly and Warech, 1994).

The empirical *JAP* data indicated very few applications of IRT or GT, which emerged during the second interval as improvements on CTT. Other than our sampling strategy, one explanation for this finding is the increased complexity of these models and another is the lag time for dissemination. Trends were noticed for several other variables. Increases occurred in self-reports, behavioral measures, and ad hoc measures, with decreases in the use of commercial measures. Behavioral and self-report measures dominated and their PUI are equivalent. Increased reporting of reliability and validity was found, but the total number of studies reporting such evidence was small. The top section of table 1.3 summarizes the PUI data for the measurement domain and figure 1.2 illustrates trends in measurement format.

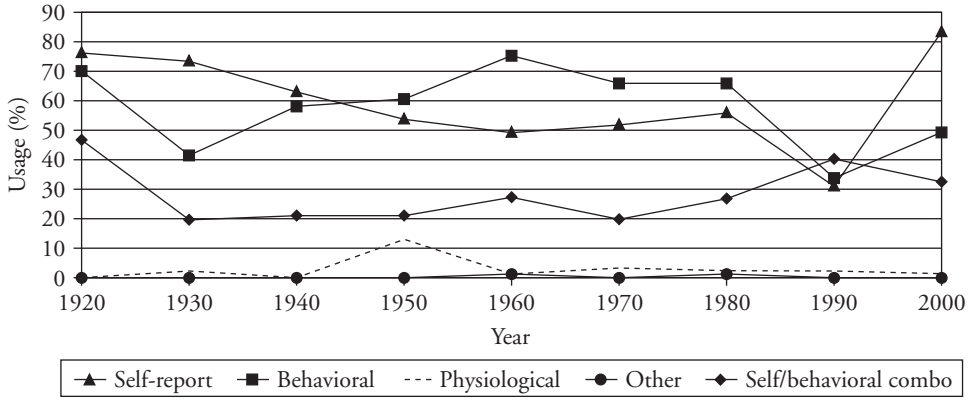

Figure 1.2 Measurement format by year

Against a backdrop of this *JAP* data, Stone-Romero's (1994) pessimism about construct evidence provided by I-O psychologists is understandable. Despite the existence of quality measures noted above, there is still a tendency to create "garden-variety" scales for studies without attending to psychometric issues. Further, despite extensive summaries of research on ability–performance relationships, there is a lack of understanding. Quantitative syntheses tell us that general ability influences supervisory evaluations through job knowledge (Hunter, 1986) – but so what? The knowledge–performance link can be elaborated if cognitive and psychometric paradigms are harnessed together. For instance, does knowledge change differ when acquired from experience rather than from training? What knowledge structures change and how do they change over time? Sternberg and associates (1999) illustrate one approach in their investigations of the role of tacit knowledge in military leadership. In another measurement arena, we observed little usage of or attention to measures that minimize reactivity (Webb, Campbell, Schwarz, Sechrest, and Grove, 1981).

Design

Design involves planning and implementing data collection, with the focus on addressing the research objectives and supporting valid inferences. The foci of design are on the degree that conditions are manipulated (e.g., passive-observational, experimental), on forming the units of analysis (e.g., within-individual, between-individual, group, cross-level), on structuring (e.g., crossing, nesting) and delivering interventions (e.g., single, multiple, repeated), on timing measurements (e.g., pre- or post-treatment), and on the setting of the research (e.g., laboratory, field). Kerlinger's (1985) Max-Min-Con dictum advises maximizing systematic variance, minimizing error variance, and controlling extraneous variance. There are few historical treatments of design (Morawski, 1988). In this section we first look at several meta-frameworks, then we examine experimental, non-experimental, and longitudinal strategies of research. Brief reviews of validity and sampling conclude the section.

Meta-frameworks range from the classic to the current. A classic includes Mill's canons of agreement, differences, agreement and differences, residues, and concomitant variation (Cowles, 1989). Major contributions were made by Cattell. One such is his data box (Cattell, 1952; see also Cronbach, 1984), which began as a cube representing persons, situations, and occasions, and which was sliced to yield different designs. Another is his decomposition of design and analysis into six parameters and derivation of basic designs (Cattell, 1988). Schaie's (1965) general developmental model separates age, period, and cohort effects and shows that only two can be estimated in any one design. Coombs's (1964) theory of data uses a fourfold classification to categorize most scalings of stimuli. These frameworks are special cases of facet theory (Guttman and Greenbaum, 1998). Their strategic benefits are in guiding elaboration of programs of research. Research on goal-setting, for example, could be elaborated by sampling widely from the data box and from the general developmental model.

A disjunction between experimental and non-experimental designs has been persistent throughout the century (Cronbach, 1957). Campbell and Stanley (1966) noted that

McCall advocated educational experimentation during the 1920s; Dehue (2000) assigns that honor to Coover at Stanford around 1913. The essence of experiments is manipulation, randomization, and control. By 1930, experimental designs evolved from the classical to the factorial, within two broad paradigms named the Wundt-Pavlov "bivariate" and the Galton-Spearman "multivariate" (Cattell, 1966). The development and usage of factorial designs in which multiple independent variables were manipulated also preceded analyses of such designs (Rucci and Tweney, 1980). A crucial concept is interaction, indicating the joint effects of multiple independent variables (Cox, 1984). Fisher's *Design of Experiments* acknowledged close relationships between design and analysis. Complex ANOVA designs became popular after World War II, and Solomon's 4-group design allowed assessment of pre-test–treatment interactions (Solomon, 1949).

Longitudinal data appeared via studies sponsored in the 1920s by the National Research Council. However, appropriate analyses of such data lagged (Lovie, 1981). Design and analysis of temporal data continues to concern methodologists (Collins and Horn, 1991; Gottman, 1995; Harris, 1963). There is a realization that the pre-post design is in most cases inadequate for the assessment of change. A current alternative, for example, synthesizes growth curve and latent variable models (Willett and Sayer, 1995).

Design of non-experimental studies was never as formalized as that of experiments. Kish's (1987) core principles of representation, randomization, and realism, which apply to all designs, are relevant. Various forms of surveys are typical instances (Dillman, 2000). Existing discussions mostly deal with sampling, item and instrument design, data collection (mail, telephone, Internet), and, recently, cognitive models of responding (Tourangeau, Rips, and Rasinski, 2000). Application of cognitive models to surveys parallels their application to test item responding.

Validity of inferences received systematic elaboration by D. T. Campbell (1957). Later, internal and external validity were advanced with a preference for internal validity (Campbell and Stanley, 1966). When elaborated into internal, statistical conclusion, external, and construct components, there was some redistribution of preferences (Cook and Campbell, 1976; Cook, Campbell, and Peracchio, 1990). The contributions included a symbolic shorthand (X, O, R, M), discussion of each type in terms of a list of "threats" and design features that could counter the threats, and quasi-experimental designs. As an example of research on external validity, Anderson, Lindsay, and Bushman (1999) reported a quantitative synthesis of research in laboratory and field settings that found "considerable correspondence" between the settings. Today, say during the current interval, social research epitomizes the experimenting society (Bickman, 2000a).

Sampling as an aspect of design supports generalizability of findings to or across populations. During the early part of the twentieth century, sampling was primarily purposive; after 1920 the importance of random sampling was realized. McNemar (1940) first reviewed sampling for psychology, and current sampling uses stratified and cluster strategies (cf. Kish, 1965). Brunswik's (1955) advocacy of representative sampling of stimuli and persons addresses generalizability and makes a link to random effects models in statistics. Dillman (2000) provides a current and thorough treatment.

Several themes emerge from this history of design. They include (1) meta-frameworks; (2) a fruitless distinction between experimental and correlational psychology; (3) the primacy of study validity; and (4) the importance of longitudinal designs.

Design developments in I-O psychology

Considering correlational design first, modal research designs prior to 1940 used convenience samples, small numbers of variables, and passive-observational strategies. Sample sizes before 1960, however, were not as small as many believe, as shown by the median N in table 1.2. Freyd (1923–4) provided a 10-step procedure for personnel selection research, while Burtt (1920) described employment research at a plant in Canada. Noteworthy features included Burtt's first developing rapport with management and measuring criteria using multiple raters, developing a battery of tests, hiring workers with lower scores to double-check the tests' predictive efficacy, and training a clerk to continue the test administration. R. L. Thorndike (1949) provided a comprehensive treatment of personnel selection based on war practice. Flanagan (1946) described a true validation of the Army Air Forces Qualification Examination and the Aircrew Classification Tests, administering these batteries, allowing all candidates who passed a physical (N = 1000) to enter training without using the other scores, and measuring short-term training success and longer-term aircrew performance.

Considering experiments within organizational psychology, the Hawthorne studies were important because they marked a shift toward studying social forces at work using "quasi"-experimental designs, and led to an early appreciation for artifacts. Viteles understood control groups in 1940 when he described their use in British experiments on vocational guidance, and he criticized Thorndike and his co-workers' use of correlational follow-up designs and general measures. Canter (1951) reviewed the use of a second control group just two years following Solomon's 4-group design. Reminiscent of the Hawthorne research, the study of large organizations via surveys and experiments emerged (Jacobsen, Kahn, Mann, and Morse, 1951). Training evaluation research emphasized experimental designs (McGehee and Thayer, 1961).

In the third interval studied, Dipboye and Flanagan (1979) disputed the truism that research in the lab is less generalizable than field research (Locke, 1986). Design of organizational surveys was discussed by Edwards, Thomas, Rosenfeld, and Booth-Kewley (1996) and by Kraut (1996). Bickman and Rog's (1998) handbook is relevant for I-O researchers with regard to design.

Recent design improvements include use of quality archival data (longitudinal, large N, quality measures). Databases now being used include Project TALENT (Austin and Hanisch, 1990), National Longitudinal Study (Dickter, Roznowski, and Harrison, 1996), and Project A (Campbell, 1990). Relatedly, researchers are moving beyond two-occasion designs due to the importance of time for the I-O field (Katzell, 1994). For example, Hofmann, Jacobs, and Baratta (1993) studied salespersons over 12 quarters to identify interindividual differences in intraindividual change (i.e., clusters of salespersons with similar patterns of change); Chan (1998) presented an integration of latent growth and longitudinal mean and covariance structure models (cf. Willett and Sayer, 1995); Tisak and Tisak (1996) presented a latent curve approach to measurement reliability and validity.

On a final note, recent work suggests that rational, maximizing design approaches may not be optimal. Martin (1982) modified a "garbage can" model that posits as key variables problems, decision-makers, choices, and solutions. Boehm (1980) described

political and nonlinear aspects of "real world" research. McGrath (1982) noted the "horns" of a dilemma, conflicting results, when evaluation is based upon multiple standards, for example trading off rigor and relevance (cf. Runkel and McGrath, 1972) or internal and external validity.

The empirical data from *JAP* for the design domain consisted of setting, strategy, and temporal structure of the design. When viewed in total, the data indicated a preponderance of field investigations (66 percent compared to 29 percent for lab studies) and very few setting combinations (1 percent). Passive-observational (PO) and field studies were consistently the most common choices across time, cross-sectional designs were overwhelmingly observed, and the populations studied were mixed between employees and students. The middle panel of table 1.3 summarizes the PUI data for the design domain, and trends in design strategy and research setting are shown in figures 1.3 and 1.4, respectively.

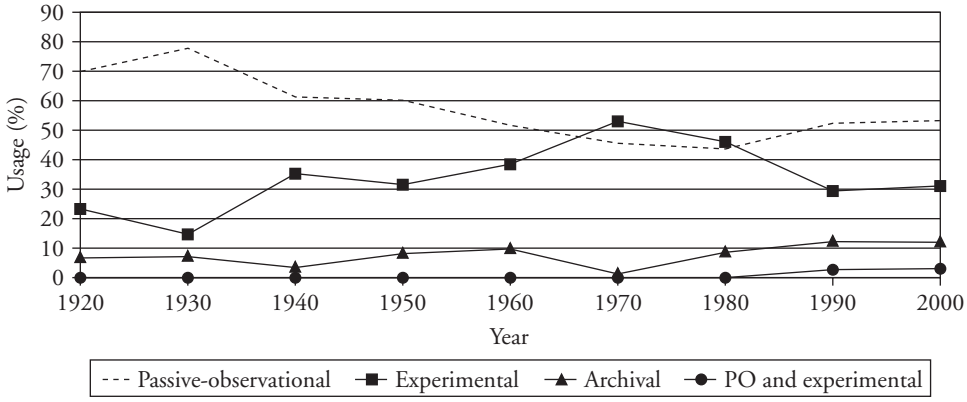

Figure 1.3 Design strategy by year

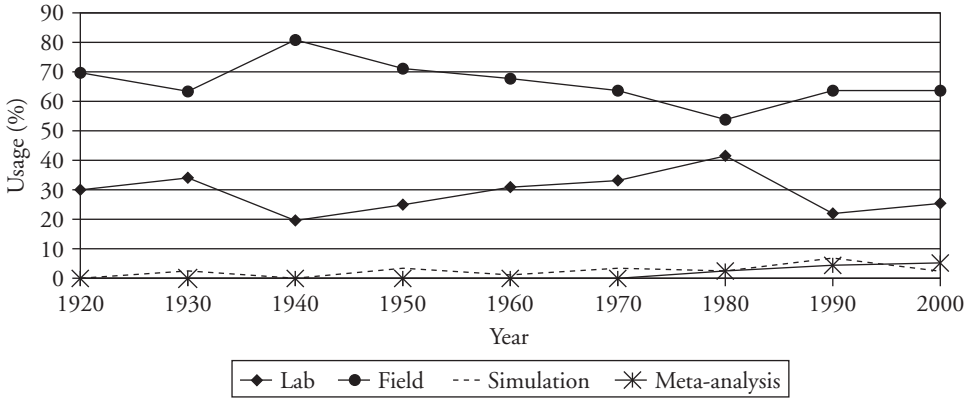

Figure 1.4 Setting usage by year

Analysis

Analysis, or statistics, consists for this review of sense-making with quantitative data. Modern statistics dates from the work of Karl Pearson, George Udny Yule, and others, as influenced by Francis Galton. Early work took place during the 1890s and progress was made rapidly after 1900. This section again examines general developments before turning to the I-O domain. The topics considered run from descriptive and inferential statistics through the diffusion of ANOVA into psychology, multivariate, popular techniques, and nonparametrics.

Descriptive statistics were well known by 1900, although it took another 75 years to bring to the fore exploratory data analysis (Tukey, 1977). Correlation and regression were well elaborated by 1920. At that time, there was a large number of unintegrated techniques pertaining to estimation, but inference was *not* well established (Kelley, 1923). Those foundations were laid by Fisher and by E. S. Pearson and Neyman between 1915 and 1935. Fisher's approach – significance testing – highlighted Type I errors, whereas E. S. Pearson and Neyman's approach – hypothesis testing – highlighted Type I and II errors. Current inferential models are a hybrid of these two approaches. Different positions and debates are thus inevitable (Chow, 1996; Harlow, Mulaik, and Steiger, 1997; Huberty and Pike, 1999; Oakes, 1986). Current work includes a taskforce report on statistical methods, written in a helpful article template style (Wilkinson and Task Force on Scientific Inference, 1999), as well as Tracey's (2000) review of null hypothesis significance testing and presentation of some ways to deinstitutionalize it.

One way to understand history is to view ANOVA as an innovation from statistics (Lovie, 1979; Rucci and Tweney, 1980). Garrett and Zubin (1943) published a crucial exposition for psychologists. Rucci and Tweney (1980) identified as the stage's initial development (1925–42) the hiatus during World War II (1942–5) and its institutionalization after the war (1945–52). Lovie (1979) identified the translation as occurring between 1934 and 1945. Fisher's 1925 textbook, *Statistical Methods for Research Workers*, contained an exposition of ANOVA and a later edition introduced ANCOVA. Subsequently, multiple comparisons evolved into planned and post hoc tests (Kirk, 1994; Ryan, 1959). By 1935 Fisher's sensitivity to the close relationship between design and analysis was incorporated in his *Design of Experiments*. The General Linear Model (GLM) is by now well understood (Cohen, 1968), but not so the Generalized Linear Model that subsumes many additional models (i.e., logistic, log-linear) through a link function (McCullagh and Nelder, 1989).

Diverse linear multivariate techniques emerged during the 1930s (Hotelling, 1936; Wilks, 1932). Other than factor analysis, which dates back to 1900, three decades elapsed before these procedures were widespread (Cattell, 1966). Most of the linear techniques are special cases of canonical correlation using various codings of the independent variables as continuous, categorical, or mixed (Knapp, 1978). Time series analysis models sequences of non-independent observations (Gottman, 1995), while event history analysis models time until event occurrence. The nonlinear techniques are more difficult to classify neatly. Techniques for studying data structure, for example, are cluster analysis and multidimensional scaling (Shepard, 1980). Bartholomew (1997), in

an historical review covering the second half of the twentieth century, reached a conclusion that software packages and bases of investigation were major problems in diffusion.

We found hierarchical linear models (HLM), covariance structure modeling (CSM), and quantitative synthesis to be currently active analytic families in terms of technical work and applications. The HLM family permits analysis at multiple levels or across multiple occasions, aiding the estimation of interindividual differences in intraindividual change (Nesselroade, 1991) as demonstrated by Hofmann et al. (1993). A second family is CSM (Bentler, 1986; MacCallum and Austin, 2000). This set of techniques, with long roots (Wright, 1934), began as an interdisciplinary synthesis of factor analysis and regression (Goldberger, 1971). Currently, general models permit the formulation of alternatives for comparison. Growth in this area is fueled by technical developments (Austin and Calderón, 1996), but researchers are the ultimate "validators" of an innovation. Jöreskog (1993) identified three strategies in model specification and evaluation: (1) strictly confirmatory, in which one a priori model is studied; (2) model generation, in which an initial model is modified until it fits well; and (3) alternative models, in which multiple a priori models are specified and then evaluated. Muthén and Curran (1997) proposed a structural equation modeling (SEM) approach to longitudinal experimental designs that explicitly incorporates power estimation. MacCallum and Austin (2000) reviewed applications, across fields of psychology, including I-O. A third and final family is quantitative synthesis, an expansion of primary and secondary analysis that uses the study or statistical test as its unit of analysis (Glass, 1976). Research syntheses appear regularly across psychology and other scientific disciplines (e.g., medicine). Models and procedures, including validity generalization, were summarized by Cooper and Hedges (1994).

Nonparametric statistics, including the χ^2 and rank correlation, were developed at the advent of modern statistics. Other developments did not occur until after 1930, however, and reached an integrated phase in the 1950s (Siegel, 1956). Clearly, the frailty of parametric statistical tests indicates the desirability of nonparametric techniques in many cases, but their usage has been infrequent within psychology, despite solid arguments (Cliff, 1996; Wilcox, 1998). A related, also underutilized, domain is the derivation of empirical sampling distributions through the bootstrap and hypothesis testing using permutation and combinatorial techniques (Arabie and Hubert, 1992; Efron, 1998; Rodgers, 1999).

Several themes emerge from this history of analysis. They include (1) the misuse surrounding statistical conclusion validity; (2) the breadth of linear models; (3) the use of multivariate methods; and (4) the improvement of analyses of change and its predictors.

Analysis developments in I-O psychology

The modal early study used descriptive and correlational statistics without inference (cf. Burtt, 1920; Terman, 1917). Inference emerged as the spread of statistical significance testing, as Hubbard, Parsa, and Luthy (1997) demonstrated by coding a random issue of each volume of *JAP* between 1917 and 1994. Hubbard and Ryan (2000) extended this research by examining a broader set of journals from 1906 to 1998. Multiple regression and partial correlation, using small numbers of predictors, were standard features of early

analyses, most of which were selection-oriented. Burtt (1926) illustrated correlation and regression in appendices. The first factor analytic study in *JAP* examined leadership ratings of female high school leaders and extracted four factors for interpretation (Flemming, 1935). That study paralleled an earlier study (Thurstone, 1932), which had analyzed Strong's correlations among 18 occupations and identified 4 vocational interest factors. Between those beginnings and 1990, roughly six decades, exploratory factor analysis was a linchpin of I-O research (Wherry and Winer, 1953). Not too many of these studies were remarkable. Fabrigar, Wegener, MacCallum, and Strahan (1999) identified flaws in exploratory factor analysis choices in *JAP* articles from 1991 to 1995. Hurley et al. (1997) presented guidance for choice between exploratory and confirmatory factor analysis (EFA/CFA), for conducting the analysis, and for interpreting the results. Their treatment could help to remedy the problems identified by Fabrigar et al. (1999), although the emergence of confirmatory factor analysis may be more helpful.

The first ANOVA published in *JAP* examined the effectiveness of deodorants and was a 3-factor design that manipulated cream, days, and persons (Jackson, Jerome, and Schoenfeld, 1942). The source table is remarkably current. We did not observe a large number of complex ANOVA analyses (e.g., fractional factorials), but ANOVA was a staple of experimental I-O research from World War II onward.

Considering multivariate techniques other than factor analysis, we located the initial applications of cluster analysis, multidimensional scaling (MDS), and canonical correlation. Cluster analysis appeared first in a study of clerical operations (Thomas, 1952), followed by MDS applied to military job performance (Schultz and Siegel, 1964) and canonical correlation applied to the Theory of Work Adjustment (Thorndike, Weiss, and Dawis, 1968). We did not find large numbers of such applications, but some I-O researchers have combined analyses in interesting ways (cf. Rodgers, 1988). McLaughlin, Carnevale, and Lim (1991) combined cluster analysis and MDS to study strategies used by negotiators. Rounds and Tracey (1993) used a synthetic approach in which they first located correlation matrices representing Holland's hexagonal model. They used those matrices to evaluate the fit of Prediger's (1982) data-ideas and people-things dimensions using confirmatory factor analysis (CFA) and MDS. Hunter (1986) first synthesized studies relating general ability to supervisory ratings through several mediators within military and civilian domains, then conducted path analysis on the aggregated correlation matrices.

Although individual levels of analysis dominate I-O psychology, some progress has been made in adding group, organization, and cross-level analysis. Group-level analyses have appeared (Kashy and Kenny, 2000; Sundstrom, McIntyre, Halfhill, and Richards, 2000). One of the important developments, both conceptually and methodologically (Katzell, 1994; Roberts, Hulin, and Rousseau, 1987), is multilevel modeling (Klein and Kozlowski, 2000; Rousseau, 1985). As noted by Hofmann (1997), HLM lends itself naturally to the study of individuals nested within departments, nested in turn within organizations.

Research synthesis – a.k.a. validity generalization (VG) – is a very useful tool that adds diversity to primary analyses (Hunter and Schmidt, 1990; Schmidt and Hunter, 1998), even though its summary statistics contain variability in their estimates of overall effect, even under sample homogeneity (Oswald and Johnson, 1998). Switzer, Paese, and Drasgow (1992) applied the bootstrap to estimate the standard error of VG statistics.

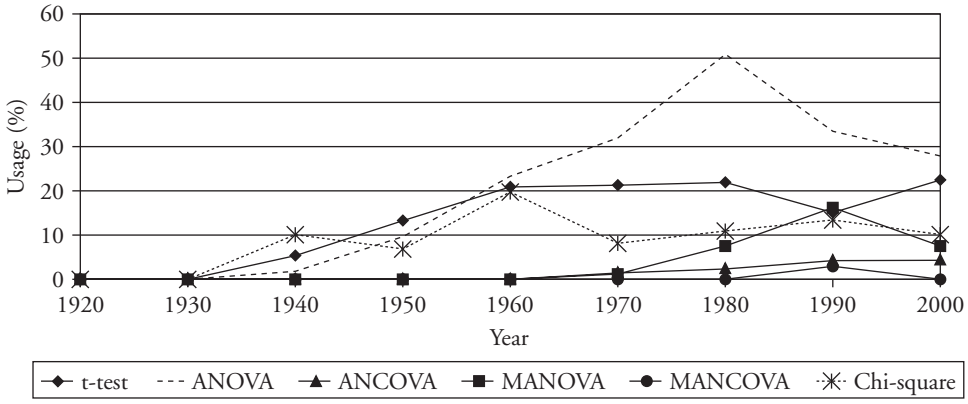

Figure 1.5 Analysis strategy by year: group comparisons

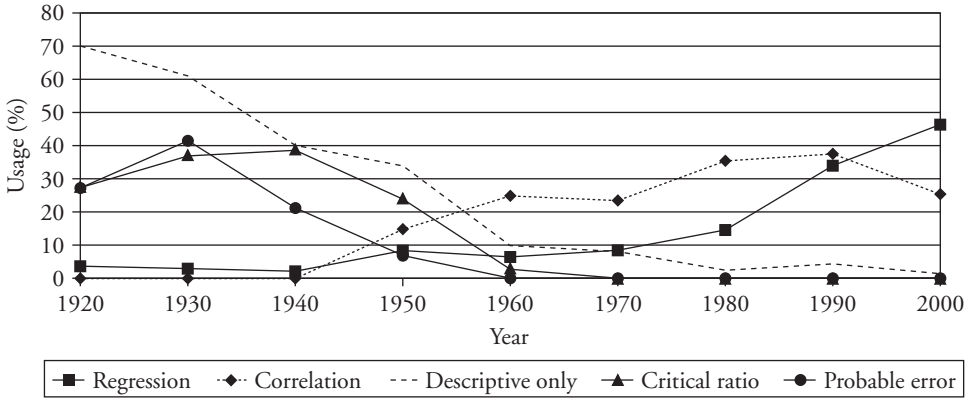

Figure 1.6 Analysis strategy by year: correlational

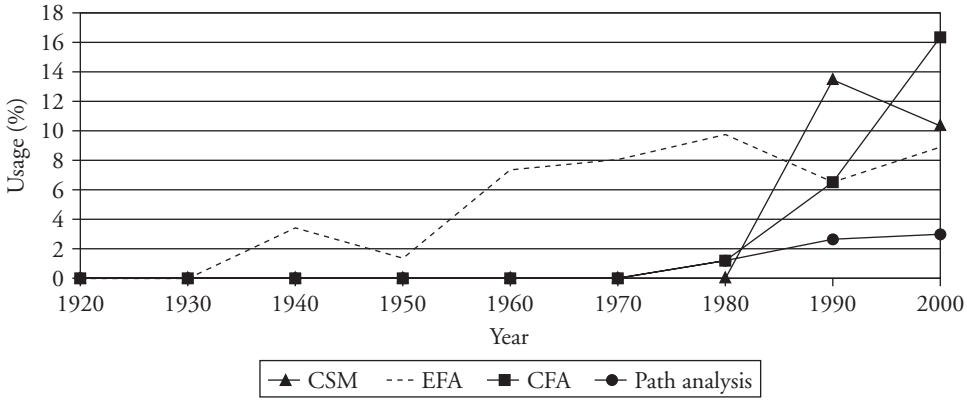

Figure 1.7 Analysis strategies by year: covariance structures

Some elevate VG to a lofty status (Schmidt, 1992), and others remain unconvinced (Bobko and Stone-Romero, 1998; James, Demaree, and Mulaik, 1986). Bobko and Stone-Romero (1998) argue comprehensively against VG as a panacea, asserting that research synthesis may merely shift some problems to the next level of aggregation. A related question pertains to the objectivity of VG (see Steiner, Lane, Dobbins, Schnur, and McConnell, 1991; Wanous, Sullivan, and Malinak, 1989). A fair conclusion is that quantitative synthesis is a crucial addition, with potential for misuse. Theoretical (Hulin, Henry, and Noon, 1990) and synthesizing (Schmitt, Gooding, Noe, and Kirsch, 1984) meta-analyses are needed.

The empirical data from *JAP* for the analysis domain revealed large changes in usage over time for some analysis techniques. The bottom section of table 1.3 contains PUI data for analysis procedures. The most notable trends are the decline in usage in the categories of critical ratio, descriptives only, and probable error, which predominated during the first interval, coupled with an increase in usage of regression and ANOVA. Starting in 1970, we observed the growth of more complex techniques, such as CSM. Figures 1.5 through 1.7 illustrate these trends for three clusters of analysis techniques: group comparisons, correlational, and covariance structures.

Discussion

What lessons can we take from the evolution of research methods within I-O? Although this chapter may have seemed to be a speedy helicopter tour of the rainforest, here we try to clarify the terrain. Finding a single summary term is difficult, but during the first interval the best characterization is *establishment*. The I-O discipline was founded in topics, roles, journals, and graduate programs (Katzell, 1991; Katzell and Austin, 1992). In measurement, I-O psychologists used abilities and vocational interests as predictors, with quantity of production as the criterion. Designs were largely correlational and cross-sectional, with small, ad hoc samples of blue-collar workers. Analyses consisted of descriptive statistics, graphs, and tables, and correlational analyses with small numbers of variables. Inferences were made with critical ratios that used the probable error in the denominator. Management was the audience. During the middle interval, 1936–68, the best characterization is *expansion*. Within the discipline, organizational psychology was born, additional training programs appeared, and professionalism emerged. In terms of research methods, most of the modern arsenal developed in measurement (i.e., construct validity, IRT), design (experimental control, sampling theory, validity threats), and analysis (ANOVA/ANCOVA, multivariate). Management remained the audience, and, to a lesser extent, unions and workers.

In the third interval, 1969–2000, one possible characterization is *eutrophication*. There has been a growth of research methods, similar to a bloom of algae, that coincides with the growth of I-O psychology (Dunnette, 1976). The complexity of research methods has changed the capability of traditional audiences to understand the level of discourse, and it seems that the audience now consists of other I-O psychologists. The peril of this approach is the gradual lessening of the relevance of the field to previous audiences.

What implications flow from the current choice set of measurement, design, and analysis procedures? All may not be well. Why? Consider that choice is becoming more difficult, misuse is increasing, and methodologists' roles are expanding (von Eye and Schuster, 2000). Quality control tools (Campion, 1993; see also this vol., ch. 22) assume motivated use by researchers, practitioners, and gate-keepers (educators, editors). Consider several factors that relate to "unfinished innovations." In measurement, a major one is incomplete adoption of IRT and GT despite cogent arguments for their use (Embretson and Hershberger, 1999). In design, incomplete innovations include designs for phenomena that are multilevel and dynamic. In analysis, a large number of incomplete innovations exist. Some of them are exploratory data analysis, robust methods for common statistical tests (Wilcox, 1998), and appropriate use of complex techniques. Others include recurrent disputes about significance testing (Harlow et al., 1997), concomitant suggestions to implement a hierarchical system that would elevate research syntheses above primary research (Schmidt, 1992, 1996), and continuing neglect of power and effect size despite well-known principles (Austin, Boyle, and Lualhati, 1998; Mone, Mueller, and Mauland, 1996). Consider further Wilcox's (1998) analysis of neglect of the effects of variance heterogeneity and non-normality upon standard analyses (r, t-test). Wilcox concluded that many discoveries have been lost through use of non-robust techniques.

Across all three domains, a clear conclusion is that computers have both facilitated and hindered research methods. One indication of facilitation is shown by the extensive publication of such "substitutes" as nomograms and abacs prior to 1960 (Scott Company, 1920; Lord, 1955). Evidence of facilitation is also seen in the ubiquity of the computer across research methods domains, from measurement via CAT (Drasgow and Olson-Buchanan, 1999), to design via optimal layout of experimental design, to analysis via software packages and "computer-intensive" resampling (Rasmussen, 1989). Additional support is evident in retrospectives (e.g., Carroll, 1987; Humphreys, 1987), these authors were explicit about expanded numbers of variables and/or occasions that could be analyzed. Hindrance occurs with mindless computer use (cf. Bartholomew, 1997). Fabrigar et al. (1999) found that a modal, yet suboptimal, set of EFA options is the default for a prominent software package.

Recent trends

What are other trends within I-O? Some recent developments are positive. One consists of investigations in which study characteristics (i.e., researcher choices) are coded and critiqued. This sort of synthesis includes reviews of significance testing (Gaither and Glorfeld, 1985), statistical power (Mone et al., 1996), sample size (Salgado, 1997), EFA (Fabrigar et al., 1999; Ford, MacCallum, and Tait, 1986), and SEM (Hulland, Chow, and Lam, 1996; MacCallum and Austin, 2000). Two examples illustrate this approach. Stone-Romero et al. (1995) coded design and data analysis features of 1,929 *JAP* articles published between 1975 and 1993. Frequencies were converted into annual PUIs for specific designs and analyses. Then these PUIs were analyzed, graphed, and evaluated. Keselman et al. (1998) demonstrated numerous problems with researcher use of ANOVA,

ANCOVA, and MANOVA analyses across a diverse set of educational and psychological journals (including *JAP*).

Another development is attention to validity. Alternatives to the Campbell–Cook framework are the validity network schema of Brinberg and McGrath (1985) and the magnitude-articulation-generality-interestingness-credibility (MAGIC) framework of Abelson (1995). One reason that validity is important to us is its unifying potential (cf. Adèr and Mellenbergh, 1999). A method for framing the unification is to note that validity pertains to supported inferences that researchers can draw based on their choices and combinations of measures, designs, and analyses. The Campbell–Cook model, applied with due diligence, requires attention to measurement of causes *and* effects (construct), to designs of research (internal, external), and to analyses (statistical conclusion validity). Messick's faceted system provides another possible unification. Several empirical studies illustrate the utility of validity frameworks for evaluating I-O choices. Cummings, Molloy, and Glen (1977) critiqued 58 work experiments using the internal–external validity distinction. Mitchell (1985) used a checklist derived from the Campbell–Cook framework and found correlational research lacking.

A third promising development, as noted above, is simulation. Ilgen and Hulin (2000) asserted that this method constitutes a "third discipline" in addition to the traditions identified by Cronbach (1957). Studies and commentary in Ilgen and Hulin (2000) address withdrawal, pay-for-performance, group decision-making, and personality inventory faking. Replications here (Axelrod, 1997) are as desirable as in any area (Neuliep, 1991).

Additional historical possibilities

Innovations could be easily studied, as examples and suggestions show. Sedlmeier and Gigerenzer (1989) hypothesized that statistical power would have increased due to Cohen's research during the 1960s (cf. Cohen, 1988). They calculated statistical power for various effects for *Journal of Abnormal Psychology* articles from 1984 and concluded that power had not increased. Another empirical approach studies current innovations as they unfold; successes and failures can be examined. Statistical conclusion validity and IRT are incomplete innovations. Case histories by innovators, for example Schaie's (1992) retrospective look at his general developmental model, constitutes a third approach. All researchers of innovation would profit from the research of Rucci and Tweney (1980), who used multiple methods to trace diffusion of ANOVA. They examined pre-1940 applications, categorized analyses across multiple journals between 1932 and 1952, identified developments in ANOVA, and reviewed textbooks and curricula. Their techniques could be applied to diffusion of neural networks, randomized response technique, or to HLM. Another approach would involve assessing psychologists across time regarding their attitudes toward innovations. Rogers (1995) groups individuals into innovators, early adopters, early majority, late majority, and laggards. A final approach exploits archival materials. Examples include I-O handbooks (1950, 1976, 1990–4), *Educational Measurement* (1951, 1972, 1989), I-O content texts (from Burtt, Viteles, and Tiffin to Landy, Schmitt and Chan, and Guion) and I-O methods texts (e.g., Schmitt

and Klimoski, 1991), guidelines for graduate education and training, and debates (e.g., Evans, 1991; Stone and Hollenbeck, 1989).

Conclusions

We have highlighted and illustrated, using a Time X Domain framework, some of the threats and opportunities that I-O researchers have faced over the years. As is true of research methodology, trade-offs are inevitable. In this chapter, we used broad strokes to paint our picture of the history of research methods in I-O psychology. This strategy precluded detailed discussions of some topics and excluded other topics altogether. We gave short shrift to issues of ethics and of theory, both of which are crucial intersections with research methods.

We acknowledge several limitations of this chapter. One is our emphasis on description, which led to an asymmetric weighting of descriptive and explanatory approaches to history. With respect to explanation, we believe that diffusion of innovation models is a crucial mechanism that can explain the evolution of methods and also lags in innovation. Another limitation was our selection of a single journal (*JAP*) and a systematic sampling plan (10th year) to provide empirical snapshots. The choice of *JAP* was dictated by its long publication history and by its prestige within the I-O field, but we recognize that additional journals and sampling plans might have yielded different results. The *Academy of Management Journal* or *Personnel Psychology*, for example, might have provided different snapshots over time.

Nevertheless, this review shows that the history of I-O research methods contains both positive and negative aspects. Greater attention to innovation will more firmly place I-O as a field on a solid footing for both research and practice. Threats to research quality will continue to exist, a perfect study will never appear, and research methods must remain a vital part of both entry-level and continuing competence in the I-O field. Some problems of misuse could be solved, we believe, by aggressive interventions in dissemination. Potential avenues include pre-convention workshops, computer mediated discussions at a distance (listservers such as RMNET and SEMNET), journals (*Organizational Research Methods, Psychological Methods*), books (e.g., the 1982 SIOP series, *Studying Organizations: Innovations in Methodology*). Drasgow and Schmitt's (2002) book on measurement and analysis, in the SIOP Frontiers series, represents an important revival of the latter approach.

Note

The authors thank Thomas Knapp, Martin Evans, Chuck Lance, David DuBois, Eric Day, Mike Coovert, Neal Schmitt, Fritz Drasgow, Rich Klimoski, and Jeff Vancouver for their comments. Especially helpful were Fred Oswald, Jim Altschuld, and Keith Widaman, whose careful and incisive critiques of earlier versions of this chapter helped substantially to improve its substance and its style. Any errancies remain with the authors.

References

Abelson, R. P. (1995). *Statistics as principled argument*. Hillsdale, NJ: Erlbaum.

Adèr, H. and Mellenbergh, G. J. (eds.) (1999). *Research methodology in the social, behavioral and life sciences*. Thousand Oaks, CA: Sage

Adkins, D. C., Primoff, E. S., McAdoo, H. L., Bridges, C. F., and Forer, B. (1947). *Construction and analysis of achievement tests*. Washington, DC: US Civil Service Commission.

Aiken, L. S., West, S. G., Sechrest, L., and Reno, R. R. (1990). Graduate training in statistics, methodology, and measurement in psychology. *American Psychologist*, 45, 721–34.

Allen, N. J. and Meyer, J. P. (1990). The measurement and antecedents of affective, continuance, and normative commitment to the organization. *Journal of Occupational Psychology*, 63, 1–18.

Anderson, C. A., Lindsay, J. J., and Bushman, B. J. (1999). Research in the psychological laboratory: Truth or triviality? *Current Directions in Psychological Science*, 8, 3–9.

Ansbacher, H. L. (1951). The history of the leaderless group discussion technique. *Psychological Bulletin*, 48, 383–91.

Arabie, P. and Hubert, L. J. (1992). Combinatorial data analysis. *Annual Review of Psychology*, 43, 169–203.

Austin, J. T. and Calderón, R. F. (1996). Theoretical and technical contributions to structural equation modeling: An updated annotated bibliography. *Structural Equation Modeling*, 3, 105–75.

Austin, J. T. and Hanisch, K. A. (1990). Occupational attainment as a function of abilities and interests: A longitudinal discriminant analysis using Project TALENT data. *Journal of Applied Psychology*, 75, 77–86.

Austin, J. T., Boyle, K., and Lualhati, J. (1998). Statistical conclusion validity for organizational science researchers: A review. *Organizational Research Methods*, 1, 164–208.

Axelrod, R. (1997). Advancing the art of simulation in the social sciences. In R. Conte, R. Hegselmann, and P. Terno (eds.), *Simulating social phenomena*. Berlin: Springer, 21–40.

Bartholomew, D. J. (1997). Fifty years of multivariate analysis. *British Journal of Mathematical and Statistical Psychology*, 50, 205–14.

Bass, B. M., Cascio, W. F., and O'Connor, E. J. (1974). Magnitude estimation of expressions of frequency and amount. *Journal of Applied Psychology*, 59, 313–20.

Baum, J. A. C. and McKelvey, B. (eds.) (1999). *Variations in organizational science: In honor of Donald T. Campbell*. Thousand Oaks, CA: Sage.

Bentler P. M. (1986). Structural modeling and *Psychometrika*: An historical perspective on growth and achievements. *Psychometrika*, 51, 35–51.

Bickman, L. (ed.) (2000a). *Validity and social experimentation*. Thousand Oaks, CA: Sage.

Bickman, L. (ed.) (2000b). *Research design*. Thousand Oaks, CA: Sage.

Bickman, L. and Rog, D. J. (eds.) (1998). *Handbook of applied social research methods*. Thousand Oaks, CA: Sage.

Binning, J. and Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology*, 74, 478–94.

Blinkhorn, S. F. (1997). Past imperfect, future conditional: Fifty years of test theory. *British Journal of Mathematical and Statistical Psychology*, 50, 175–85.

Bobko, P. and Stone-Romero, E. F. (1998). Meta-analysis may be another useful research tool, but it is not a panacea. *Research in Personnel and Human Resources Management*, 16, 359–97.

Bock, R. D. (1997). Some history of item response theory. *Educational Measurement: Issues and Practice*, 16, 21–33.

Boehm, V. R. (1980). Research in the "real world" – A conceptual model. *Personnel Psychology*, 33, 495–503.

Brennan, R. L. (1997). A perspective on the history of generalizability theory. *Educational Measurement: Issues and Practice*, 16, 14–20.

Brinberg, D. and McGrath, J. E. (1985). *Validity and the research process*. Newbury Park, CA: Sage.

Brown, C. W. and Ghiselli, E. E. (1955). *Scientific method in psychology*. New York: McGraw-Hill.

Browne, M. W. (2000). Psychometrics. *Journal of the American Statistical Association*, 95, 661–5.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, 62, 193–217.

Burtt, H. E. (1920). Employment psychology in the rubber industry. *Journal of Applied Psychology*, 4, 1–20.

Burtt, H. E. (1926). *Principles of employment psychology*. New York: Harper.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54, 297–312.

Campbell D. T. and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.

Campbell, D. T. and Stanley, J. (1966). Experimental and quasi-experimental designs for research. Chicago, IL: Rand-McNally.

Campbell, J. P. (1990). An overview of the army selection and classification project (Project A). *Personnel Psychology*, 43, 231–9.

Campbell, J. P. (1996). Group differences and personnel decisions: Validity, fairness, and affirmative action. *Journal of Vocational Behavior*, 49, 122–58.

Campion, M. A. (1993) Article review checklist: A criterion checklist for reviewing research articles in applied psychology. *Personnel Psychology*, 46, 705–18.

Canter, R. R. (1951). The use of extended control-group designs in human relations studies. *Psychological Bulletin*, 48, 340–7.

Carroll, J. B. (1987). Measurement and educational psychology: Beginnings and repercussions. In J. A. Glover and R. R. Ronning (eds.), *Historical foundations of educational psychology*. New York: Plenum, 89–106.

Cascio, W. F., Outtz, J., Zedeck, S., and Goldstein, I. (1991). Statistical implications of six methods of test score use in personnel selection. *Human Performance*, 4, 233–64.

Cattell, R. B. (1952). The three basic factor-analytic research designs – their interrelations and derivatives. *Psychological Bulletin*, 49, 499–520.

Cattell, R. B. (ed.) (1966). *Handbook of multivariate experimental psychology* (1st edn). Chicago: Rand-McNally.

Cattell, R. B. (1988). The principles of experimental design and analysis in relation to theory building. In J. Nesselroade and R. B. Cattell (eds.), *Handbook of multivariate experimental psychology* (2nd edn). New York: Plenum Press, 21–67.

Chan, D. (1998). The conceptualization and analysis of change over time: An integrative approach incorporating longitudinal mean and covariance structures analysis (LMACS) and multiple indicator latent growth modeling (MLGM). *Organizational Research Methods*, 1, 421–83.

Chow, S. L. (1996). *Statistical significance: Rationale, validity, and utility*. Thousand Oaks, CA: Sage.

Cliff, N. R. (1996). *Ordinal methods for behavioral data analysis*. Mahwah, NJ: Erlbaum.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin*, 70, 426–43.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd edn). Hillsdale, NJ: Erlbaum.

Collins, L. M. and Horn, J. L. (eds.) (1991). *Best methods for the analysis of change*. Washington, DC: American Psychological Association.

Committee on Classification of Personnel (1919). *The personnel system of the United States Army*. Washington, DC: Department of War.

Cook, T. D. and Campbell, D. T. (1976). The design and conduct of quasi-experiments and true experiments in field settings. In M. D. Dunnette (ed.), *Handbook of industrial and organizational psychology*. Chicago, IL: Rand McNally, 223–326.

Cook, T. D., Campbell, D. T., and Peracchio, L. (1990). Quasiexperimentation. In M. D. Dunnette and L. A. Hough (eds.), *Handbook of industrial and organizational psychology* (2nd edn, vol. 1). Palo Alto, CA: Consulting Psychologists' Press, 491–576.

Cook, J. D., Hepworth, S. J., Wall, T. D., and Warr, P. B. (1981). *The experience of work*. New York: Academic Press.

Coombs, C. H. (1964). *A theory of data*. New York: Wiley.

Cooper, H. M. and Hedges, L. V. (eds.) (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.

Cowles, M. (1989). *Statistics in psychology: An historical perspective*. Hillsdale, NJ: Erlbaum.

Cox, D. R. (1984). Interaction. *International Statistical Review*, 52, 1–31.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, 12, 671–83.

Cronbach, L. J. (1975). Five decades of controversy over mental testing. *American Psychologist*, 30, 1–14.

Cronbach, L. J. (1984). A research worker's treasure chest. *Multivariate Behavioral Research*, 19, 223–40.

Cronbach, L. J. (1991). Methodological studies: A personal retrospective. In R. E. Snow and D. E. Wiley (eds.), *Improving inquiry in social science*. Hillsdale, NJ: Erlbaum, 385–400.

Cronbach, L. J. and Gleser, G. (1957). *Psychological tests and personnel decisions*. Urbana: University of Illinois Press.

Cronbach, L. J. and Meehl, P. C. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.

Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Cummings, T. G., Molloy, E. S., and Glen, R. (1977). A methodological critique of fifty-eight selected work experiments. *Human Relations*, 30, 675–703.

Dawis, R. V. (1987). Scale construction. *Journal of Counseling Psychology*, 34, 481–9.

Dawis, R. V. and Lofquist, L. H. (1993). Rejoinder: From TWA to PEC. *Journal of Vocational Behavior*, 43, 113–21.

Dehue, T. (2000). From deception trials to control reagents: The introduction of the control group about a century ago. *American Psychologist*, 55, 264–68.

Dickter, D., Roznowski, M. A., and Harrison, D. A. (1996). Temporal tempering: An event history analysis of the process of voluntary turnover. *Journal of Applied Psychology*, 81, 705–16.

Dillman, D. (2000). *Mail and Internet surveys* (2nd edn). New York: Wiley.

Dipboye, R. L. and Flanagan, M. F. (1979). Research settings in industrial-organizational psychology: Are findings in the field more generalizable than in the laboratory? *American Psychologist*, 34, 141–50.

Drasgow, F. (1982a). Choice of test model for appropriateness measurement. *Applied Psychological Measurement*, 6, 297–308.

Drasgow, F. (1982b). Biased test items and differential validity. *Psychological Bulletin*, 92, 526–31.

Drasgow, F. and Hulin, C. L. (1990). Item response theory. In M. D. Dunnette and L. A. Hough (eds.), *Handbook of industrial and organizational psychology* (2nd edn, vol. 1). Palo Alto, CA: Consulting Psychologists' Press, 577–636.

Drasgow, F. and Olson-Buchanan, J. B. (eds.) (1999). *Innovations in computerized assessment*. Mahwah, NJ: Erlbaum.

Dunnette, M. D. (1976). Toward fusion. In M. D. Dunnette (ed.), *Handbook of industrial and organizational psychology*. Chicago: Rand-McNally, 1–12.

Edwards, J. E., Thomas, M. D., Rosenfeld, P., and Booth-Kewley, S. (1996). *How to conduct organizational surveys: A step-by-step guide*. Thousand Oaks, CA: Sage.

Efron, B. (1998). R. A. Fisher in the 21st century. *Statistical Science*, 13, 95–122.

Embretson, S. E. and Hershberger, S. L. (eds.) (1999). *The new rules of measurement*. Mahwah, NJ: Erlbaum.

Evans, M. G. (1991). The problem of analyzing multiplicative composites: Interactions revisited. *American Psychologist*, 46, 6–15.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., and Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4, 272–99.

Ferguson, L. W. (1950). The LOMA merit rating scales. *Personnel Psychology*, 3, 193–216.

Fine, S. A. (1955). Functional job analysis. *Personnel Administration and Industrial Relations*, 2, 1–16.

Flanagan, J. C. (1946). The experimental evaluation of a selection procedure. *Educational and Psychological Measurement*, 6, 445–66.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin*, 51, 327–58.

Fleishman, E. A. and Mumford, M. D. (1991). Evaluating classifications of job behavior: A construct validation of the ability requirement scales. *Personnel Psychology*, 44, 523–75.

Flemming, E. G. (1935). A factor analysis of personality of high school leaders. *Journal of Applied Psychology*, 19, 596–605.

Ford, J. K., MacCallum, R., and Tait, M. (1986). The application of exploratory factor analysis in applied psychology: A critical review and analysis. *Personnel Psychology*, 39, 291–314.

Freyd, M. (1923–24). Measurement in vocational selection: An outline of research procedure. *Journal of Personnel Research*, 2, 215–49, 268–84, 377–85.

Gaither, N. and Glorfeld, L. (1985). An evaluation of the use of tests of significance in organizational behavior research. *Academy of Management Review*, 10, 787–93.

Garrett, H. E. and Zubin, J. (1943). The analysis of variance in psychological research. *Psychological Bulletin*, 40, 233–67.

Ghiselli, E. E., Campbell, J. P., and Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. San Francisco, CA: Freeman.

Glass, G. (1976). Primary, secondary, and meta analysis of research. *Educational Researcher*, 5, 3–8.

Goffin, R. D. and Helmes, E. (eds.) (2000). *Problems and solutions in human assessment*. Boston, MA: Kluwer.

Goldberger, A. S. (1971). Econometrics and psychometrics: A survey of communalities. *Psychometrika*, 36, 83–107.

Goldstein, H. and Wood, R. (1989). Five decades of item response modeling. *British Journal of Mathematical and Statistical Psychology*, 42, 139–167.

Gottfredson, L. (ed.) (1986). The g factor in employment. *Journal of Vocational Behavior*, 29, (special issue), 293–450.

Gottfredson, L. and Sharf, J. (eds.) (1988). Fairness in employment testing. *Journal of Vocational Behavior*, 33, (Special issue), 225–477.

Gottman, J. M. (ed.) (1995). *The analysis of change*. Mahwah, NJ: Erlbaum.

Guilford, J. P. (1936). *Psychometric methods*. New York: Appleton-Century-Crofts.

Guion, R. M. (1977). Content validity – The source of my discontent. *Applied Psychological Measurement*, 1, 1–10.

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions*. Mahwah, NJ: Erlbaum.
Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.
Guttman, R. and Greenbaum, C. W. (1998). Facet theory: Its development and current status. *European Psychologist*, 3, 13–36.
Hakel, M. (ed.) (1998). *Beyond multiple choice: Evaluating alternatives to traditional testing for selection*. Mahwah, NJ: Erlbaum.
Hanson, F. A. (1993) *Testing testing*. Berkeley, CA: University of California Press.
Harlow, L. L., Mulaik, S. A., and Steiger, J. H. (eds.) (1997). *What if there were no significance tests*? Mahwah, NJ: Erlbaum.
Harrell, T. W. (1992). Some history of the Army General Classification Test. *Journal of Applied Psychology*, 77, 875–8.
Harris, C. W. (ed.) (1963). *Problems in measuring change*. Madison: University of Wisconsin Press.
Hartigan, J. A. and Wigdor, A. K. (eds.) (1989). *Fairness in employment testing*. Washington, DC: National Academy Press.
Harvey, R. J. (1993). *The development of the Common Metric Questionnaire*. (www.pstc.com)
Herrnstein, R. J. and Murray, C. E. (1994). *The bell curve: Intelligence and class structure in the United States*. New York: Free Press.
Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management*, 21, 967–88.
Hofmann, D. A. (1997). An overview of the logic and rationale of hierarchical linear models. *Journal of Management*, 23, 723–44.
Hofmann, D. A., Jacobs, R., and Baratta, J. (1993). Dynamic criteria and the measurement of change. *Journal of Applied Psychology*, 78, 194–204.
Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28, 321–77.
Howard, A. and Bray, D. W. (1988). *Managerial lives in transition*. New York: Guilford Press.
Hubbard, R. and Ryan, P. A. (2000). The historical growth of statistical significance testing in psychology – and its future prospects. *Educational and Psychological Measurement*, 60, 661–81.
Hubbard, R., Parsa, R. A., and Luthy, M. R. (1997). The spread of statistical significance testing in psychology. *Theory and Psychology*, 7, 545–54.
Huberty, C. J. and Pike, C. J. (1999). On some history regarding statistical testing. *Advances in Social Science Methodology*, 5, 1–22.
Hulin, C. L., Drasgow, F., and Parsons, C. (1983). *Item response theory: Applications to psychological measurement*. Homewood, IL: Dow Jones-Irwin.
Hulin, C. L., Henry, R. A., and Noon, S. L. (1990). Adding a dimension: Time as a factor in the generalizability of predictive relationships. *Psychological Bulletin*, 107, 328–40.
Hull, C. L. (1928). *Aptitude testing*. Yonkers, NY: World Books.
Hulland, J., Chow, Y. H., and Lam, S. (1996). Use of causal models in marketing research: A review. *International Journal of Research in Marketing*, 13, 181–97.
Humphreys, L. G. (1987). Quantitative methodology: Then, now, and the future. In J. A Glover and R. R. Ronning (eds.), *Historical foundations of educational psychology*. New York: Plenum, 403–14.
Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior*, 29, 340–62.
Hunter, J. E. and Schmidt, F. L. (1990). *Methods of meta-analysis*. Newbury Park, CA: Sage.
Hurley, A., Schriesheim, C. A., Brannick, M. T., Seers, A., Vandenberg, R. J., and Williams, L. J. (1997). Exploratory and confirmatory factor analysis: Guidelines, issues, and alternatives. *Journal of Organizational Behavior*, 18, 667–83.

Ilgen, D. R. and Hulin, C. L. (2000). *Computational modeling of behavior in organizations: The third scientific discipline*. Washington, DC: American Psychological Association.

Ironson, G., Smith, P. C., Brannick, M. T., Gibson, W. M., and Paul, K. B. (1989). Construction of a "Job in General" scale: A comparison of global, composite, and specific measures. *Journal of Applied Psychology*, 74, 193–200.

Jackson, T. A., Jerome, E. A., and Schoenfeld, N. (1942). Experimental and statistical analysis of the effectiveness of deodorant creams. *Journal of Applied Psychology*, 26, 308–15.

Jacobsen, E., Kahn, R., Mann, F. C., and Morse, N. C. (eds.) (1951). Human relations research in large organizations. *Journal of Social Issues*, 7(3) (Special issue).

James, L. R. (1973). Criterion models and construct validity. *Psychological Bulletin*, 80, 75–83.

James, L. R. (1998). Measurement of personality via conditional reasoning. *Organizational Research Methods*, 1, 131–63.

James, L. R., Demaree, R. G., and Mulaik, S. A. (1986). A note on validity generalization procedures. *Journal of Applied Psychology*, 71, 440–50.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.

Johns, G. (1993). Constraints on the adoption of psychology-based personnel practices: Lessons from organizational innovation. *Personnel Psychology*, 46, 569–92.

Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen and J. S. Long (eds.), *Testing structural equation models*. Newbury Park, CA: Sage, 294–316.

Kashy, D. A. and Kenny, D. (2000). The analysis of data from dyads and groups. In H. Reis and C. Judd (eds.), *Handbook of research methods in social and personality psychology*. Cambridge, England: Cambridge University Press, 567–93.

Katzell, R. A. (1991). History of early I-O doctoral programs. *The Industrial-Organizational Psychologist*, 28(4), 51.

Katzell, R. A. (1994). Contemporary meta-trends in industrial and organizational psychology. In H. C. Triandis, M. D. Dunnette, and L. M. Hough (eds.), *Handbook of industrial and organizational psychology* (2nd edn, vol. 4). Palo Alto, CA: Consulting Psychologists Press, 1–93.

Katzell, R. A. and Austin, J. T. (1992). From then to now: The development of industrial-organizational psychology in the United States. *Journal of Applied Psychology*, 77, 803–35.

Kelley, T. L. (1923). *Statistical method*. New York: Macmillan.

Kerlinger, F. N. (1985). *Foundations of behavioral research* (3rd edn). New York: Holt, Rinehart, and Winston.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue. B., Kowalchuk. R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., and Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, 68, 350–86.

Kirk, R. E. (1994). Choosing a multiple-comparison procedure. *Advances in Social Science Methodology*, 3, 77–121.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Kish, L. (1987). *Statistical design for research*. New York: Wiley.

Klein, K. J. and Kozlowski, S. W. J. (eds.) (2000). *Multilevel theory, research, and methods in organizations*. San Francisco, CA: Jossey-Bass.

Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance-testing system. *Psychological Bulletin*, 85, 410–16.

Kraut, A. I. (ed.) (1996). *Organizational surveys: Tools for assessment and change*. San Francisco, CA: Jossey-Bass.

Kuder, G. F. and Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2, 151–60.

Kyllonen, P. C. (1997). Smart testing. In R. F. Dillon (ed.), *Handbook on testing*. Westport, CT: Greenwood Press, 347–71.

Lawler, E. E., III, Nadler, D. A., and Cammann, C. (eds.) (1980). *Organizational assessment*. New York: Wiley.

Lewis, C. (1986). Test theory and *Psychometrika*: The past twenty-five years. *Psychometrika*, 51, 11–22.

Locke, E. A. (ed.) (1986). *Generalizing from laboratory to field settings*. Lexington, MA: Lexington Press.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3, 635–94.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph No. 7*.

Lord, F. M. (1955). Nomographs for computing multiple correlation coefficients. *Journal of the American Statistical Association*, 50, 1073–7.

Lord, F. M. and Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lovie, A. D. (1979). The analysis of variance in experimental psychology: 1934–1945. *British Journal of Mathematical and Statistical Psychology*, 32, 151–78.

Lovie, A. D. (1981). On the early history of ANOVA in the analysis of repeated measures in psychology. *British Journal of Mathematical and Statistical Psychology*, 34, 1–15.

Lowman, R. (1996). What every psychologist should know about assessment. *Psychological Assessment*, 7, (Special section), 339–68.

Lumsden, J. (1976). Test theory. *Annual Review of Psychology*, 27, 251–80.

MacCallum, R. A. (1998). Commentary on quantitative methods in I-O research. *The Industrial-Organizational Psychologist*, 35(4), 18–30.

MacCallum, R. A. and Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, 51, 201–26.

McCormick, E. J., Jeanneret, P. R., and Meachem, R. C. (1969). *The development and background of the Position Analysis Questionnaire (PAQ)*. West Lafayette, IN: Occupational Research Center.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models* (2nd edn). New York: Chapman and Hall.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

McGehee, W. and Thayer, P. W. (1961). *Training in business and industry*. New York: Wiley.

McGrath, J. E. (1982). Dilemmatics: The study of research choices and dilemmas. In J. E. McGrath, J. Martin, and R. A. Kukla (eds.), *Judgment calls in research*. Beverly Hills, CA: Sage, 69–102.

McLaughlin, M. E., Carnevale, P., and Lim, R. G. (1991). Professional mediators' judgments of mediation tactics: Multidimensional scaling and cluster analyses. *Journal of Applied Psychology*, 76, 465–72.

McNemar, Q. (1940). Sampling in psychological research. *Psychological Bulletin*, 37, 331–65.

Martin, J. (1982). A garbage can model of the research process. In J. E. McGrath, J. Martin, and R. A. Kulka (eds.), *Judgment calls in research*. Beverly Hills, CA: Sage, 17–39.

Meijer, R. R. and Nering, M. L. (eds.) (1999). Computerized adaptive testing. *Applied Psychological Measurement*, 23(3), (special issue).

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' response performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741–9.

Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin*, 100, 398–407.

Michell, J. (1999). *Measurement in psychology: Critical history of a methodological concept*. Cambridge, England: Cambridge University Press.

Mitchell, T. R. (1985). An evaluation of the validity of correlational research conducted in organizations. *Academy of Management Review*, 10, 192–205.

Mone, M. A., Mueller, G. C., and Mauland, W. (1996). The perceptions and usage of statistical power in applied psychology and management research. *Personnel Psychology*, 49, 103–20.

Morawski, J. G. (ed.) (1988). *The rise of experimentation in American psychology*. New Haven, CT: Yale University Press.

Mosier, C. I. (1940). Psychophysics and mental test theory. I. Fundamental postulates and elementary theorems. *Psychological Review*, 47, 355–366.

Mowday, R. T., Steers, R. M., and Porter, L. W. (1979). The measurement of organizational commitment. *Journal of Vocational Behavior*, 14, 224–247.

Muthén, B. O. and Curran, P. J. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods*, 2, 371–402.

Nesselroade, J. (1991). Interindividual differences in intraindividual change. In L. M. Collins and J. L. Horn (eds.), Best methods for the analysis of change. Washington, DC: American Psychological Association, 92–105.

Neuliep, J. (ed.) (1991). *Replication research in the social sciences*. Newbury Park, CA: Sage.

Oakes, M. R. (1986). *Statistical inference: A commentary for the social and behavioural sciences*. Chichester, England: Wiley.

Olson-Buchanan, J. B., Drasgow, F., Moberg, P. J., Mead, A. P., Keenan, P., and Donovan, M. A. (1998). Interactive video assessment of conflict resolution skills. *Personnel Psychology*, 51, 1–24.

OSS Assessment Staff. (1948). *Assessment of men: Selection of personnel for the Office of Strategic Services*. New York: Rinehart.

Oswald, F. L. and Johnson, J. W. (1998). On the robustness, bias, and stability of statistics from meta-analysis of correlation coefficients: Some initial Monte Carlo findings. *Journal of Applied Psychology*, 83, 164–78.

Owen, D. B. (ed.) (1976). *On the history of statistics and probability*. New York: Marcel Dekker.

Pedhazur, E. and Schmelkin, L. (1991). *Measurement, design, and analysis*. Hillsdale, NJ: Erlbaum.

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., and Fleishman, E. A. (eds.) (1999). *An occupational information system for the 21st century: The development of O*Net*. Washington, DC: American Psychological Association.

Prediger, D. J. (1982). Dimensions underlying Holland's hexagon: Missing link between interests and occupations? *Journal of Vocational Behavior*, 21, 259–87.

Price, J. L. (1997). Handbook of organizational measurement. *International Journal of Manpower*, 18, 301–558.

Rasmussen, J. L. (1989). Computer-intensive correlational analysis: Bootstrap and approximate randomization techniques. *British Journal of Mathematical and Statistical Psychology*, 42, 103–11

Reilly, R. R. and Warech, M. A. (1994). The validity and fairness of alternatives to cognitive tests. In H. Wing and B. R. Gifford (eds.), *Policy issues in employment testing*. Boston: Kluwer, 131–224.

Roberts, K. H., Hulin, C. L., and Rousseau, D. (1987). *Developing an interdisciplinary science of organizations*. San Francisco, CA: Jossey-Bass.

Rodgers, J. L. (1988). Structural models of the American Psychological Association in 1986: A taxonomy for reorganization. *American Psychologist*, 43, 372–82.

Rodgers, J. L. (1999). The bootstrap, the jackknife, and the randomization test: A sampling taxonomy. *Multivariate Behavioral Research*, 34, 441–56.

Rogers, E. M. (1995). *Diffusion of innovations* (4th edn). New York: Free Press.

Rounds, J. and Tracey, T. J. (1993). Prediger's dimensional representation of Holland's RIASEC circumplex. *Journal of Applied Psychology*, 78, 875–90.

Rousseau, D. (1985). Issues of level in organizational research: Multilevel and cross-level perspectives. *Research in Organizational Behavior*, 7, 1–37.

Rucci, A. J. and Tweney, R. D. (1980). Analysis of variance and the "second discipline" of scientific psychology: A historical account. *Psychological Bulletin*, 87, 166–84.

Runkel, P. J. and McGrath, J. E. (1972). *Research on human behavior: A systematic guide to method*. New York: Holt, Rinehart, and Winston.

Ryan, T. (1959). Multiple comparisons in psychological research. *Psychological Bulletin*, 56, 26–47.

Sackett, P. R. and Larson, J. R., Jr. (1990). Research strategies and tactics in industrial and organizational psychology. In M. D. Dunnette and L. M. Hough (eds.), *Handbook of industrial and organizational psychology* (2nd edn, vol. 1), Palo Alto, CA: Consulting Psychologists Press, 419–89.

Sackett, P. R. and Wilk, S. L. (1994). Within-group norming and other forms of score adjustment in pre-employment testing. *American Psychologist*, 49, 932–54.

Salgado, J. F. (1997). Sample size in validity studies of personnel selection. *Journal of Occupational and Organizational Psychology*, 71, 161–4.

Sands, W. A., Waters, B. K., and McBride, J. R. (eds.) (1997). *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.

Schaie, K. W. (1965). A general model for the study of developmental problems. *Psychological Bulletin*, 64, 92–107.

Schaie, K. W. (1992). The impact of methodological changes in gerontology. *International Journal of Aging and Human Development*, 35, 19–29.

Schmidt, F. L. (1992). What do data really mean? *American Psychologist*, 47, 1173–81.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods*, 1, 115–29.

Schmidt, F. L. and Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124, 262–74.

Schmitt, N. and Klimoski, R. J. (1991). *Research methods in human resources management*. Cincinnati, OH: South-Western.

Schmitt, N. and Landy, F. J. (1993). The concept of validity. In W. Borman and N. Schmitt (eds.), *Personnel selection in organizations*. San Francisco, CA: Jossey-Bass, 275–309.

Schmitt, N., Gooding, R. Z., Noe, R. A., and Kirsch, M. (1984). Meta-analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology*, 37, 407–22.

Schultz, D. G. and Siegel, A. I. (1964). The analysis of job performance by multi-dimensional scaling techniques. *Journal of Applied Psychology*, 48, 329–35.

Schwab, D. T. (1980). Construct validity in organizational behavior. *Research In Organizational Behavior*, 2, 3–43.

Scott, W. D. (1917). A fourth method of checking results in vocational selection. *Journal of Applied Psychology*, 1, 61–6.

Scott Company. (1920). Tables to facilitate the computation of coefficients of correlation by rank differences method. *Journal of Applied Psychology*, 4, 115–25.

Sedlmeier, P. and Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105, 309–16.

Shavelson, R. J., Webb, N. M., and Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, 44, 922–32.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, 210, 390–8.

Siegel, S. (1956). *Nonparametric statistics*. New York: McGraw-Hill.

Smith, P. C., Kendall, L., and Hulin, C. L. (1969). *The measurement of satisfaction in work and retirement*. Chicago: Rand-McNally.

Solomon, R. (1949). An extension of control group design. *Psychological Bulletin*, 46, 137–50.

Steiner, D. D., Lane, I. M., Dobbins, G. H., Schnur, A., and McConnell, S. (1991). A review of meta-analyses in organizational behavior and human resources management: An empirical assessment. *Educational and Psychological Measurement*, 51, 609–27.

Sternberg, R. J., et al. (1999). *Tacit knowledge in the workplace* (Technical Report 1093). Alexandria, VA: Army Research Institute.

Sternberg, R. J. and Wagner, R. K. (1993). The g-ocentric view of intelligence and job performance is wrong. *Current Directions in Psychological Science*, 1, 1–5.

Stone, E. F. and Hollenbeck, J. R. (1989). Clarifying some controversial issues surrounding statistical procedures for detecting moderator variables: Empirical evidence and related matters. *Journal of Applied Psychology*, 74, 3–10.

Stone-Romero, E. F. (1994). Construct validity issues in organizational behavior research. In J. Greenberg (ed.), *Organizational behavior: The state of the science*. Hillsdale, NJ: Erlbaum, 155–79.

Stone-Romero, E. F., Weaver, A. E., and Glenar, J. L. (1995). Trends in research design and data analytic strategies in organizational research. *Journal of Management*, 21, 141–57.

Sundstrom, E., McIntyre, M., Halfhill, T., and Richards, H. (2000). Work groups: From the Hawthorne studies to work teams of the 1990s and beyond. *Group Dynamics*, 4, 44–67.

Switzer, F. S., III, Paese, P. W., and Drasgow, F. (1992). Bootstrap estimates of standard errors in validity generalization. *Journal of Applied Psychology*, 77, 123–29.

Terman, L. M. (1917). A trial of mental and pedagogical tests in a civil service examination for policemen and firemen. *Journal of Applied Psychology*, 1, 17–29.

Thomas, L. L. (1952). A cluster analysis of office operations. *Journal of Applied Psychology*, 36, 238–42.

Thorndike, R. L. (1949). *Personnel selection*. New York: Wiley.

Thorndike, R. M., Weiss, D. J., and Dawis, R. V. (1968). Multivariate relationships between a measure of vocational interests and a measure of vocational needs. *Journal of Applied Psychology*, 52, 491–6.

Thurstone, L. L. (1919a). Mental tests for prospective telegraphers: A study of the diagnostic value of mental tests for predicting ability to learn telegraphy. *Journal of Applied Psychology*, 3, 110–17.

Thurstone, L. L. (1919b). A standardized test for office clerks. *Journal of Applied Psychology*, 3, 248–51.

Thurstone, L. L. (1931). *Reliability and validity of tests (mimeo)*. Chicago, IL: University of Chicago.

Thurstone, L. L. (1931–32). A multiple factor study of vocational interests. *Personnel Journal*, 10, 198–205.

Tisak, J. and Tisak, M. (1996). Longitudinal models of reliability and validity: A latent curve approach. *Applied Psychological Measurement*, 20, 275–88.

Toops, H. A. (1944). The criterion. *Educational and Psychological Measurement*, 4, 271–97.

Tourangeau, R., Rips, L. J., and Rasinski, K. (eds.) (2000). *The psychology of survey response*. New York: Cambridge University Press.

Tracey, T. J. G. (2000). Issues in the analysis and interpretation of quantitative data: Deinstitutionalization of the null hypothesis test. In S. D. Brown and R. W. Lent (eds.), *Handbook of counseling psychology* (3rd. edn). New York: Wiley.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice*, 16, 8–14.

Tryon, R. C. (1957). Reliability and behavior domain validity: Reformulation and historical critique. *Psychological Bulletin*, 54, 229–49.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Vandenberg, R. J. and Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3, 4–69.

Van der Linden, W. J. and Hambleton, R. K. (eds.) (1997). *Handbook of modern item response theory*. Berlin: Springer.

Viteles, M. S. (1932). *Industrial psychology*. New York: Norton.

Von Eye, A. and Schuster, C. (2000). The road to freedom: Quantitative developmental methodology in the third millennium. *International Journal of Behavioral Development*, 24, 35–43.

Von Mayrhauser, R. T. (1992). The mental testing community and validity: A prehistory. *American Psychologist*, 47, 244–53.

Wanous, J., Sullivan, S. E., and Malinak, J. (1989). The role of judgment calls in meta-analysis. *Journal of Applied Psychology*, 74, 259–64.

Webb, E. J., Campbell, D. T., Schwarz, R. J., Sechrest, L., and Grove, J. B. (1981). *Nonreactive measures in the social sciences*. Dallas, TX: Houghton-Mifflin.

Wherry, R. J. and Winer, B. J. (1953). A method for factoring large numbers of items. *Psychometrika*, 18, 161–79.

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist*, 53, 300–14.

Wilkinson, L. and Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594–604.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, 24, 471–93.

Willett J. B. and Sayer A. G. (1995). Cross-domain analyses of change over time: Combining growth modeling and covariance structure analysis. In G. A. Marcoulides, R. E. Schumacker (eds.), *Advanced structural equation modeling: Issues and techniques*. Mahwah, NJ: Erlbaum, 125–57.

Wilpert, B. (1997). One hundred years of work and organizational psychology. In R. Fuller, P. L. Noonan-Walsh, and P. McGinley (eds.), *A century of psychology*. London: Routledge, 192–206.

Wing, H. C. and Gifford, B. R. (eds.) (1994). *Policy issues in employment testing*. Boston: Kluwer.

Wright, B. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice*, 16, 33–45, 52.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics*, 5, 161–215.

Yerkes, R. M. (ed.) (1921). Psychological examining in the United States Army. *Memoirs of the National Academy of Sciences*, 15.