American Board of Police and Public Safety

Core Scientific Knowledge for Specialists in Police and Public Safety Psychology[1]

Edition 1.0 Assessment Domain
October 2014

---

[1] Prepared by Michael Cuttler, Ph.D., ABPP; Paul Detrick, Ph.D., ABPP; Jay Supnick, Ph.D., ABPP; Joanne Brewster, Ph.D., ABPP and Casey Stewart, Psy.D., ABPP and approved by the ABPPSP Board of Directors in October 2014

Table of Contents

Series Introduction

This document has been developed by the American Academy of Police & Public Safety Psychology, and is designed to guide candidates for board certification as they prepare for written and oral specialty examinations. This document is intended to provide a subject matter "refresher" for those who are familiar with the basic underlying material as a result of their professional education. As such, it is recommended that this material be viewed as a study tool, i.e. a set of suggestions and reminders of key scientific concepts that inform evidence-based practice in police and public safety psychology. However, it is also equally important to remember that the discussions that follow are not meant to be a substitute for primary scientific source material. Similarly, the references and/or suggested readings cited in these discussions are meant to be representative of the subject matter at hand, i.e. suggestions of a place to start to pursue further investigation and/or to refresh your knowledge of basic concepts, but should by no means be considered a comprehensive review of all literature relevant to these topics nor a comprehensive reading list for ABPPSP examinations.

This edition of the Core Scientific Knowledge document is focused on only one of the four domains of police and public safety psychology (PPSP) - Assessment. Subsequent editions will focus on other domains of practice (Intervention, Operations, Organizational Consultation), However, some of the material presented herein is likely to be relevant to multiple practice domains. Inasmuch as PPSP specialists are expected to be conversant across all domains of practice, it is recommended that all candidates for board certification in police and public safety psychology review all information presented under each domain heading in preparation for the Board certification process.

The primary content areas for discussion within this edition (Assessment) are:

Theory, history, and systems:

Specialists should be thoroughly conversant with the primary theories, history, and systems of practice; scholarly work; practice guidelines; major issues; and controversies relevant to the primary domain of practice. These include :

1. Definition and history of the discipline, parameters of the evidence base in this domain.
2. Important scholarly works.
3. Theoretical Issues.
    i. Primary methodological positions (behavioral vs. psychodynamic therapy, industrial/organizational vs. clinical employment assessment).

ii. Current theoretical issues and controversies (e.g., on the job performance measurement vs. simulation, mandatory vs. voluntary professional intervention, CISD etc.).

4. Practice standards.

Research design and statistics:

5. Characteristics, principles, and limitations of common research designs and data gathering techniques.
6. Fundamental statistical techniques (descriptive, inferential, correlational, factorial).
7. Factors affecting the accuracy of conclusions of psychological research and/or predictions made by psychological tests and evaluation processes.

Instrument construction, prediction, therapeutic prognostics, outcome measurement, and/or program evaluation:

Specialists are expected to understand and be conversant in theory, methods, and critical concepts relevant to prediction and/or behavioral measurement of individual/group differences and outcomes (i.e., they should be able to identify, describe, and critically evaluate common instruments, procedures, and processes). For example, assessment psychologists should be familiar with the design and technical characteristics of the assessment instruments they use in their evaluations as well as the assumptions, scoring rubrics, and other procedures they use to reach conclusions and recommendations, and the limitations of each.

Using the information in this document when preparing for the Professional Self-Study Statement (PSSS) in support of ABPPSP Board Certification

The information in this core knowledge source document has been designed to assist candidates for board certification in formulating their response to section D (Scientific Base) of the Professional Self-Study Statement (PSSS). Details of these requirements may be found in the most recent version of the ABPPSP Examination Manual (see www.abppsp.org).

The focus of this area of the PSSS is the description of the research evidence that informs assessment practice in PPSP. It is expected that candidates for board certification in PPSP will respond to this area by referencing, specifically, the totality of evidence that informs their assessment practices.

When describing scientific evidence, candidates should gather, analyze, critically evaluate, report, and synthetize the totality of evidence with reference to demonstration of specialty level competence. The ABPPSP Board of Directors regards specialist-level competency in Police & Public Safety Psychology as conduct and decision-making that reflects the totality of evidence and conforms to the national standard of practice. Several kinds of professional authority and sources of scientific evidence are relied upon when defining standards of practice (Heilbrun, DeMatteo, Marczyk, & Goldstein, 2008), in descending order of importance:

1. The APA Ethics Code and pertinent regulations, laws, and case law, which apply to all psychologists.

2. Professional practice guidelines and clinical practice guidelines published by the American Psychological Association linked to scientifically-derived evidence following a rigorous, formal review process with input from multiple stakeholders in professional psychology, and regulatory enforcement guidance (e.g., EEOC guidance for enforcement of ADA and GINA).

3. Publications that articulate broad principles and are developed using multiple sources of authority.

4. An overall description of research and practice as offered in the literature, through a national survey of views or practices, or a meta-analysis of empirical research.

5. Professional practice guidelines prepared through consensus among practitioners (e.g., IACP-PPSS guidelines).

6. Manuals for psychological tests or other assessment instruments that are carefully attentive to reliability and validity and help a reader assess the quality of the instrument and place it in the broader context of specialty practice.

7. The systematic review of what recognized scholars and practitioners in the field have written and taught regarding the elements that comprise competent practice.

8. A single study describing a survey, or offering an empirical description, of some particular aspect of specialty practice.

Candidates for specialist in all domains of PPSP practice should describe the scientific evidence supporting the effectiveness of their practice activities to include the totality of evidence considered when choosing the instruments, protocols or procedures utilized. Typically this evidence will include peer-reviewed, parochial, and clinical references. Similarly, candidates should describe the validity and/or outcome measures cited in the peer-reviewed and/or parochial literature that support their own practice.

Furthermore, candidates for specialist practice in the assessment domain should describe the design and/or psychometric characteristics of the instruments and scoring protocols used, as well as the evidence that supports their use (peer-reviewed, parochial, and clinical), including the interpretive strategy applied in practice and the evidence that supports use of this strategy.

Representative Topical Reading List for ABPPSP Board Certification (Science)

As noted earlier, the purpose of this core knowledge document is to provide a "summary refresher" for those who are familiar with the basic underlying material as a result of their professional education. As such, it is recommended that this material be viewed as a study tool(i.e., a set of suggestions and reminders of key scientific concepts that inform evidenced-based practice in police and public safety psychology).

The following section presents a list of readings that are representative of the subject matter presented in the rest of this document. They were drawn from the larger reference list that appears at the conclusion of this document and were selected by ABPPSP SME's and the authors of this document on the basis of their direct relevance to the topic areas under which they fall. In combination with the topical explanations covered in this document, these readings may be helpful as a refresher/re-orientation when preparing for the science components of ABPPSP specialty examinations.

NOTE: *this list of readings **DOES NOT** represent an exhaustive set of topics and/or data used to construct any component of the specialty examination process. (**i.e., topics/items covered in the suggested readings may or may not be represented in the specialty examination**)*

Representative Topical Reading List for ABPPSP Board Certification (Science)

As noted earlier, the purpose of this core knowledge document is to provide a "summary refresher" for those who are familiar with the basic underlying material as a result of their professional education. As such, it is recommended that this material be viewed as a study tool(i.e., a set of suggestions and reminders of key scientific concepts that inform evidenced-based practice in police and public safety psychology).

The following section presents a list of readings that are representative of the subject matter presented in the rest of this document. They were drawn from the larger reference list that appears at the conclusion of this document and were selected by ABPPSP SME's and the authors of this document on the basis of their direct relevance to the topic areas under which they fall. In combination with the topical explanations covered in this document, these readings may be helpful as a refresher/re-orientation when preparing for the science components of ABPPSP specialty examinations.

History and Systems in Police Psychology: The Assessment Domain

EEOC – Uniform Guidelines on Employee Selection Procedures (1978), § 1607.4 para D

International Association of Chiefs of Police. (2009). Psychological Fitness-for-Duty Evaluation Guidelines. Arlington, VA: International Association of Chiefs of Police.

Kitaeff, J. History of Police Psychology. In Kitaeff J. (Ed.). (2011). *Handbook of Police Psychology*. New York, NY: Routledge.

Lefkowitz, J. (1977). Industrial-organizational psychology and the police. *American Psychologist*, 32(5), 346-364.

Title VII of the Civil Rights act (Civil Rights Act of 1964, Title VII, Pub. L. 88-352 (78 Stat. 241))

Weiss, P.A. & Inwald, R. (2010). A brief history of personality assessment in police psychology. In P.A. Weiss (Ed.) *Personality Assessment in Police Psychology: A 21st Century Perspective*. 5-28 Springfield, IL: Charles C. Thomas.

Wiggins, J.S. (2003). *Paradigms of Personality Assessment.* New York, NY: Guilford.

Clinical vs. Actuarial or Statistical Prediction Models

Cuttler, M. J. (2011). Pre-employment screening of police officers: Integrating actuarial prediction models with practice. In J. Kitaeff, Handbook of Police Psychology (pp. 135-163). New York, NY: Routledge.

Faust, D. & Ahern, D. S. (2012). Clinical judgment and prediction. In D. Faust (Ed), Coping with Psychiatric and Psychological Testimony (pp. 147-208) (6th ed.) . Oxford: Oxford University Press.

Faust, D., Ziskin, M (2012). Coping with Psychiatric and Psychological Testimony (6^{th} Edition) Oxford: Oxford University Press.Fine, S. A. & Cronshaw, S. F. (1999). *Functional job analysis: A foundation for human resources management*. Mahwah, NJ: Erlbaum.

Grove, W.M., & Meehl, P.E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. Psychology, Public Policy, and Law, 2, 293–323.

Grove, W.M., Zald, D.H., Lebow, B.S., Snitz, B.E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. Psychological Assessment, 12, 19-30.

Meehl, P.E. (1954). Clinical vs. statistical prediction. Minneapolis, MN: University of Minnesota

Westen, D. & Weinberger, J. (2004). When clinical description becomes statistical prediction. American Psychologist, 59, 595-613.

Pre conditional offer of employment and Bifurcated pre-employment evaluations

Berner, John G (2010) Pre-offer personality testing in the selection of entry-level California peace officers (technical report), A report to the California Commission on Peace Officer Standards and Training; http://lib.post.ca.gov/Publications/technical_report.pdf

Cuttler, M. J. (2011). Pre-employment screening of police officers: Integrating actuarial prediction models with practice. In J. Kitaeff, Handbook of Police Psychology (pp. 135-163). New York, NY: Routledge.

Enforcement guidance: Pre-employment disability-related inquiries and medical examinations under the Americans With Disabilities Act of 1990. (1995). Equal Employment Opportunity Commission, ADA Division, Office of Legal Counsel. Washington, DC.

Stewart, Casey O. (2008) The Validity of the California Psychological Inventory in the Prediction of Police Officer Applicants Suitability for Employment . School of Professional Psychology. Paper 156 .http://commons.pacificu.edu/spp/156

Additional Issues Related to Psychological Assessments of Police and Public Safety Personnel

Corey, D., & Borum, R. (2013). Forensic assessment for high---risk occupations. In R. K. Otto (Ed.), Forensic psychology (pp. 246---270). Vol. 11 in I. B. Weiner (Editor---in---Chief). Handbook of psychology (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Corey, D.M. (2011) Principles of Fitness-for-Duty Evaluations for Police Psychologists, in J. Kitaeff, (Ed.) Handbook of Police Psychology (pp 263-293). New York, NY: Routledge.

Corey, D. M. (2012). Core Legal Knowledge in Police & Public Safety Psychology. Paper presented at the American Board of Professional Psychology Summer Workshop Series, Boston, MA, July 11, 2012.

Research Design and Statistics

Beins, B. C. (2013). Research methods: A tool for life (3rd ed.). NY: Pearson.

Finn, S. E. (2009). Incorporating base rate information in daily clinical decision making. In J. N. Butcher (Ed.), Oxford handbook of personality assessment (pp. 140-149). New York, NY, Oxford University Press.

Hunter, J. E., & Schmidt, F. L. (1990). Methods of meta-analysis: Correcting error and bias in research findings. Newbury Park, CA: Sage.

Keller, D. K. (2006). The tao of statistics: A path to understanding (with no math). Thousand Oaks, CA: Sage.

Lipsey, M. W., & Wilson, D. B. (2001). Practical meta-analysis. Applied social research methods series (Vol. 49). Thousand Oaks, CA: Sage Publications.

Ones, D. S. & Viswesvaran, C. (2003). Job-specific applicant pools and national norms for personality scales: Implications for range-restriction corrections in validation research. Journal of Applied Psychology, 88, 570-577.

Common uses, techniques, instruments for measurement, and/or outcome evaluation within the assessment domain

Aamodt, M. G. (2004). Research in law enforcement selection. Boca Raton, FL: Brown Walker Publishing.

Corey, D.M. (2011) Principles of Fitness-for-Duty Evaluations for Police Psychologists, in J. Kitaeff, (Ed.) Handbook of Police Psychology (pp 263-293). New York, NY: Routledge.

Cuttler, M. J. (2011). Preemployment screening of police officers: Integrating actuarial prediction models with practice. In J. Kitaeff, Handbook of Police Psychology (pp. 135-163). New York, NY: Routledge.

Weiss, P.A. & Inwald, R. (2010). A brief history of personality assessment in police psychology. In P.A. Weiss (Ed.) Personality Assessment in Police Psychology: A 21st Century Perspective. 5-28 Springfield, IL: Charles C. Thomas.

Reliability and Test Construction (Standard error of measurement, internal consistency measures, factors effecting reliability of outcomes and measures)

Allen, M.; Yen W. (1979). Introduction to Measurement Theory. Monterey, CA: Brooks/Cole.

Cascio, W. F. (1991). Applied psychology in personnel management (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334.

Nunnally, J. C. (1978). Psychometric theory (2nd ed.). New York, NY: McGraw-Hill.

Test Validity

AERA, APA & NCME (1999). Standards for educational and psychological testing. Washington, DC: AERA, APA, and NCME.

Cascio, W. F. (1991). Applied psychology in personnel management (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Civil Rights Act of 1964, Title VII, Pub. L. 88-352 (78 Stat. 241).

Fine, S. A. & Cronshaw, S. F. (1999). Functional job analysis: A foundation for human resources management. Mahwah, NJ: Erlbaum.

Muchinsky, P. M. (2012). Criteria; Standards for Decision Making (chapter 3) and Predictors, Psychological Assessment (chapter 4) in Muchinsky, Paul M Psychology Applied to Work. Summerfield, NC: Hypergraphic Press, Inc.

Principles for validation and use of personnel selection procedures (4th Ed.). (2003). Society for Industrial and Organizational Psychology. Washington, DC: American Psychological Association.

Robinson, M. (2012). What is Job Analysis? Institute of Work Psychology. Retrieved from. http://esrccoi.group.shef.ac.uk/pdf/whatis/job_analysis.pdf

Other important factors affecting utility of measurement instruments/processes (e.g. disparate impact)

Biddle, D. (2005). Adverse Impact and Test Validation: A practitioner's Guide to Valid and
Defensible Employment Testing. London England: Gower.

EEOC – Uniform Guidelines On Employee Selection Procedures (1978), § 1607.4 para D

Roth, P. L., Bobko, P. L., & Switzer, F. S. III. (2006). Modeling the behavior of the 4/5ths rule for determining adverse impact: Reasons for caution. Journal of Applied Psychology, 91, 507-522.

History and Systems in Police Psychology: The Assessment Domain

The purpose of this section is to review systems of practice and major theoretical issues relevant to assessment in police and public safety psychology. This will include: 1) a brief definition and description of the scope of police psychology practice in the area of psychological assessment; 2) a history of police psychology assessment; and 3) current theoretical issues (e.g., clinical vs. actuarial prediction; bifurcated models of assessment).

Kitaeff (2011) defines police psychology as, "The application of psychological principles and methods to law enforcement. This growing area includes topics such as screening and hiring of police officers, conducting screening for special assignments (e.g., SWAT), fitness-for-duty evaluations, investigations, hostage negotiations training and consultation, and stress counseling, among others." The focus of the current discussion is primarily psychological assessment of police officer candidates; however, assessment in police and public safety psychology is not limited solely to employment selection. In this regard, Aumiller at al. (2007) identified four domains of practice in police and public safety psychology, including at least 20 specific assessment-related activities. Among these were fitness-for-duty, threat assessment, promotional assessments, 360 evaluations, and measurement of the outcome of interventions.

There are several good book chapters available that describe the history of police psychology (cf. Blau, 1994; Kitaeff, 2011; Reese, 1995; Weiss and Inwald, 2010). The reader is referred to these chapters for a comprehensive treatment of the subject. Since this document is being written for psychologists considering certification as specialists in police and public safety psychology, it will only contain a discussion of the essential points that candidates for the American Board of Professional Psychology (ABPP) Certification in Police & Public Safety Psychology should have as a foundation for understanding history and systems as it relates to the assessment domain in police and public safety psychology.

Dating the beginning of a field is a difficult task, even for professional historians, since there are often many milestones along the way to mark a number of actual beginnings or innovations that come together to define a field. These early beginnings are the "primordial soup" from which the current field of police and public safety psychology developed. Thus, it is important to know that the pioneers of modern psychology began applying their theories and methods to the world of policing in the early part of the twentieth century.

It is generally acknowledged that Lewis Terman, the developer of the Stanford-Binet Intelligence Test, was the first psychologist to use psychological testing in the selection of police officers when he used a modified form of the test with police candidates at the San Jose Police Department in California in 1916. Terman published his results in 1917

and concluded that intelligence testing could be potentially valuable as a selection tool (Weiss and Inwald, 2010, p.6). Thurstone (1922) used the Army Alpha Intelligence test to assess officers in the Detroit Police Department and found that patrolmen scored higher than their commanding officers. A study of personality using the Rorschach Inkblot Test, scored using the Klopfer (1946) scoring method, was done at the New York City Police Department in 1950.. Correlations were found between test scores, job satisfaction, and motivation for promotion (Kates, 1950). Blau (1994) reported that before 1950 there were some early attempts to use "personality instruments" in the selection of police officers. These efforts for the first half of the century were important but generally reflected the efforts of academic or experimental psychologists venturing into the police psychology world, where they applied developing methodologies and assessment tools. The late 1960's and early 1970's were a time of rapid and sometimes violent change in society with much social unrest and turmoil fueled by the Civil Rights Movement and Vietnam War. It was also a time of change and development in the selection of police officers. In 1967, the President's Commission on Law Enforcement and the Administration of Justice recommended that psychological testing be used to screen police officer candidates. These recommendations were incorporated into the Omnibus Crime Control and Safe Streets Act of 1968. The U.S. National Advisory Commission on Civil Disorders (1968) recommended the elimination of police officers whose performance was impacted by personal prejudice. Funding for research on psychological testing of police officers also came in 1968 from the Law Enforcement Assistance Administration. By 1973, the National Commission on Criminal Justice Standards and Goals included recommendations for "behavioral science resources" in police departments. Clearly, these were transformative times, and efforts by the federal government fueled the further development of police psychology and the psychological assessment of police candidates.

The federal government continued to provide support for the development of police psychology through the 1970's. Most notable in this regard were several conferences sponsored by the FBI that focused on psychological assessment of police officer candidates, where police psychologists and researchers came together to share information.[2] The FBI conferences were the impetus for the formation of several professional organizations for police psychology. Initially, the Council on Police Psychology (COPP) and the Academy of Police Psychologists (APP) continued communication among psychologists practicing in police settings and started developing standards and guidelines for the provision of services (Blau, 1994). Eventually other groups were formed, including the Society for Police and Criminal Psychology (SPCP; 1975), the Police and Public Safety Section of Division 18 of the American Psychological

[2] Two of these conferences were the National Working Conference on the Selection of Law Enforcement Officers in 1976 and the World Conference on Police Psychology in 1985 at the FBI Academy in Quantico, VA.).

Association (APA PPSS; 1983), and the Police Psychological Services Section of the International Association of Chiefs of Police (IACP PPSS; October 25, 1984).

Around the same time, there were developments in the legal and regulatory world, particularly Title VII of the Civil Rights act (Civil Rights Act of 1964, Title VII, Pub. L. 88-352 (78 Stat. 241)) and the Equal Employment Opportunity Commission's Uniform Guidelines on Employee Selection, which were published in 1978. The Uniform Guidelines spelled out the necessary validation requirements of all selection methods with appropriate criterion validity, thus increasing motivation among police psychologists to scientifically validate their test batteries and selection procedures. Furthermore, requirements to avoid adverse impact were also referenced in the Civil Rights Act of 1991 and clarified by the EEOC (2007). Legal and regulatory influences have continued to influence the development and use of new assessment methods in police psychology, most notably the Americans with Disabilities Act of 1990 (and its amendment, the Americans with Disabilities Act Amendments Act [ADAAA, 2008]) and the Civil Rights Act of 1991, and most recently, the Genetic Information Non Discrimination Act of 2008 (GINA).

For the past 20 years, the legal and regulatory influences on pre-employment psychological evaluation have been profound in the area of police psychological assessment, because they have forced psychologists to use scientific methods to defend their approach to assessment and the selection of tools. There has been increasing sophistication with regard to job analysis, meta-analytic studies of prediction outcome, and use of new selection tools, such as bio-data.

In response to legal and regulatory changes, professional groups began to publish guidelines for psychologists. In 1987, the Society of Industrial and Organizational Psychology (SIOP) published the first edition of the Principles for the Validation and Use of Personnel Selection Procedures (Hereafter, the Principles). Over the years, there have been successive revisions to the Principles, with the most recent version, the 4th Edition, having been published in 2003. The IACP began developing the Pre-employment Psychological Evaluation Guidelines, first adopted in 1986, and the Psychological Fitness-for-Duty Guidelines, first adopted in 1991; and subsequently revised every five years since then, the most recent FFDE versions were revised and ratified in 2009 and 2013, respectively.

The Society for Police and Criminal Psychology (SCPC) was the first organization to develop an advanced certification (diplomate) in police psychology, requiring that an applicant demonstrate advanced knowledge in the field by passing a written and oral examination. In 2007, a joint effort between the Police Psychological Services Section of the International Association of Chiefs of Police (IACP-PPSS) and SPCP defined in detail the domains and proficiencies of police psychology (Aumiller, et al., 2007). In 2008, these two organizations, along with the American Psychological Association's Division 18 Police Psychological Services Section, petitioned the American Psychological

Association's Commission for the Recognition of Specialties and Proficiencies in Professional Psychology (CRSPPP) to recognize Police & Public Safety Psychology as a proficiency in professional psychology (Weiss and Inwald, 2010). In 2010, Police and Public Safety Psychology was recognized as an affiliated specialty board by the American Board of Professional Psychology (ABPP) and began to award ABPP Board Certification in 2011. Finally, in 2013 Police and Public Safety Psychology was recognized as a full specialty by the Commission for the Recognition of Specialties and Proficiencies in Professional Psychology Council (CRSPPP) of the American Psychological Association. The increasing recognition of police and public safety psychology as a specialty in professional psychology continues to help raise the level of professionalism, communicate standards of practice, and enhance the competency of psychologists practicing police and public safety psychology, in general, and providing assessment services for police and public safety agencies, in particular.

Clinical vs. Actuarial or Statistical Prediction Methods

All data obtained in the course of a psychological assessment should be analyzed, interpreted and reported in a systematic manner with analysis and conclusions informed through review of scientific evidence. There are two commonly recognized methods of accomplishing this task; approaches grounded in clinical judgment and those based on actuarial (also called statistical and/or mechanical) interpretation. The clinical method has been defined as development of a set of impressionistic subjective conclusions, informed by a combination of the scientific knowledge, experience and clinical judgment by the examiner. The actuarial method of prediction method of prediction is "a formal method, employs an equation, a formula, a graph, or an actuarial table to arrive at a probability, or expected value, of some outcome" (Grove & Meehl, 1996).

In any event, it is critical that police psychologists make sure that their efforts conform to the scientific evidence base and that the rationale and interpretative protocols engaged are clearly articulated. Doing so requires an understanding and appreciation of both clinical and actuarial interpretation strategies. The debate about clinical vs. actuarial prediction is one of the most persistent controversies in psychology. It has been ongoing since Meehl first published his book "Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence" (Meehl, 1954). Yet, in spite of the fact that the data in support of the actuarial model have been both plentiful and incontrovertible, with research to the contrary being absent and/or fundamentally flawed, many practicing psychologists continue to rely on clinical judgment when interpreting assessment findings. (cf. Dawes, Faust, & Meehl, 1989; Goldberg, 1970; Grove, Zald, Hallberg, Lebow, Snitz, & Nelson, 2000; Vrieze & Grove, 2009).

Burgess (1928) is credited as the first researcher to show that a crude actuarial formula for combining data was superior to clinical judgment in predicting parole outcome

(Grove & Meehl, 1996). Since that time, the data have been overwhelmingly consistent in study after study: finding results of actuarial prediction equals to or better than clinical prediction by a significant margins (cf. Dawes, Faust, & Meehl, 1989; Goldberg, 1970; Grove, Zald, Hallberg, Lebow, Snitz, & Nelson, 2000; Vrieze & Grove, 2009; Faust & Ahern, 2012). [3]

Dawes (2005) notes that in some measure reliance on clinical judgment approaches may be due to a common misunderstanding among psychologists about the nature of actuarial assessment, particularly as it relates to the earlier term "statistical prediction". Actually, the term statistical prediction is rarely used within this context today; i.e. the way that the data are aggregated does not have to actually involve complex statistical calculations, per se. Rather, the fundamental facet of actuarial prediction is the pre hoc establishment of interpretative rules and protocols which are applied in a systematic manner. The application of these rules might involve statistical procedures, such as a multiple regression/ discriminate equations etc. , but may also involve a set of simple decision rules, algorithms etc. which have been developed and validated through both experience and scientific endeavor. (Cuttler, 2011). [4]

It should be noted that there are also models of prediction combining clinical judgment and actuarial methods, however, none of these methods have been found to be superior to actuarial prediction alone (Dawes, Faust & Meehl, 1989; Grove, Zald, Lebow, Snitz, & Nelson, 2000). Hanson and Morton-Bourgon (2009) conducted a meta-analysis of alternative decision methods (actuarial procedures, clinical judgment, and structured professional judgment). They also examined a "hybrid method" where summary scores were obtained based on a set of items or factors used to evaluate risk (clinical and actuarial mix). The results indicated, in order of accuracy: actuarial procedures, hybrid method, structured professional judgment, clinical judgment. Thus, all else being held equal, "… the use of formal or structured data collection as opposed to non-standardized or unstructured approaches enhances accuracy" (Faust & Ahern, 2012, p. 158).

Similarly, research supporting the notion of incremental validity derived from other than structured interview techniques is also somewhat sparse. Sawyer (1966) discussed the usefulness of interview data when structured in the collection process and processed through actuarial mechanisms. More recently research has also shown that structured

---

[3] See Grove, W.M., & Meehl, P.E. (1996). This is considered to be a seminal work comparing 136 empirical studies and addressing common anti actuarial arguments.

[4] Grove and Meehl (1996) also use the term "mechanical prediction" which is a more accurate expression of the fundamental process underlying actuarial assessment.

aggregation of interview data primarily improves prediction because it also reduces error variance (Westen and Weinberger, 2004). Similarly Faust & Ziskin, (2012) discuss how clinical observation and description can be aggregated to support actuarial prediction.

Pre-conditional offer and Bifurcated Pre-employment Psychological Evaluations

Prior to 1990, pre-employment psychological evaluations (PEPEs) for public safety candidates were a relatively straightforward endeavor. PEPEs were administered to candidates prior to being hired, and they generally contained measures of psychopathology, such as an MMPI-2, PAI, or IPI, and measures of normal behavior, such as the CPI or 16PF. In 1990, the Americans with Disabilities Act (ADA) was introduced into law with major implications and consequences for PEPEs.

The Americans with Disabilities Act of 1990 was enacted to prevent discrimination toward individuals with medical disabilities, whether physical or mental. Title I of the ADA has to do with disabilities in employment situations. In order to prevent discrimination based on a disability, Title I limits an employer's ability to make disability-related inquiries or to require medical examinations prior to an offer of employment. It stated that "an employer may ask disability-related questions and require medical examinations of an applicant only after the applicant has been given a conditional job offer" (Enforcement Guidance: Preemployment Disability-Related Questions and Medical Examinations, 1995) .

In order to comply with these new regulations, most police psychologists started to administer PEPEs after a conditional offer of employment was given to a candidate by the hiring agency. While this initially seemed to satisfy the ADA requirements, further case law clarified the ADA requirement of providing a "bona-fide" conditional offer of employment before the administration of a medical examination. This meant that all non-medical information that could reasonably be expected to be collected was already considered. In 2005, the case of *Lionel v. American Airlines*, in the 9th Circuit United States Court of Appeals, ruled that American Airlines did not provide a bona-fide conditional offer of employment because the appellants were administered a medical examination before the background investigation was completed, and did not meet the business necessity exception from the ADA mandate to gather all reasonable non-medical hiring information before a medical examination is conducted.

Lionel v. American Airlines highlighted a problem with administering PEPEs after a conditional offer of employment if the PEPEs measured normal personality traits, since these traits should have been taken into consideration in the initial vetting for the conditional offer of employment. The solution to this problem was to "bifurcate" the

PEPEs evaluation into two parts: the first measuring normal personality characteristics and the second measuring medically-related information, (i.e., psychopathology). The Enforcement Guidelines specifically state, "Psychological tests that measure personality traits such as honesty, preferences, and habits are not considered medical examinations." In this regard a modest number of research efforts focused on documentation of predictive validity of non-medical tests has emerged (Stewart 2008; Berner 2010).

The Bifurcated PEPEs Model (BPM) divides the psychological evaluation process into Pre-offer and Post-Offer assessment in order to comply with the ADA. Nevertheless, the BPM and practice guidelines are still evolving (Cuttler, 2011). One controversy of the BPM is whether a pre-offer evaluation should contain an interview performed by a clinical psychologist. Presumably, such an interview could possibly identify medical problems, causing a violation of ADA. Indeed, the Enforcement Guidelines contain the following indicators to consider when determining if a test or procedure is a medical examination: (1) whether the test is administered by a health care professional; and (2) whether the test is interpreted by a health care professional. Clearly, an interview by a clinical psychologist meets these two criteria. Nevertheless, some assert that a clinical psychologist could interview a candidate at the pre-offer stage if the questions asked were non-medical. Even though the EEOC has endorsed this position (Rennert[5], personal communication), the controversy remains, and the issue remains untested in a court of law. Without making a recommendation for or against interviews by a psychologist with clinical training at the pre-offer stage, anyone considering performing interviews at this stage should consider the potential risks and benefits of such practice to all parties.

Additional Issues Related to Psychological Assessments of Police and Public Safety Personnel

As discussed in the introduction to this section, there are a number of different kinds of assessments that are conducted on police and public safety personnel or applicants. Pre-employment psychological evaluations were already discussed briefly, and although this is the most common type of assessment activity that police and public safety psychologists engage in, fitness-for-duty examinations, threat assessments, and promotional evaluations are a few other examples of the many assessment services that police and public safety psychologists provide. A review of each of the various

[5] Sharon Rennert , Senior Attorney Advisor, Americans With Disabilities Act Division, Office Of Legal Counsel, U.S. Equal Employment Opportunity Commission

types of assessments is beyond the scope of this paper, but there are many sources that discuss each in detail (see Aumiller et al., 2007; Blau, 1994; Kitaeff, 2011; Kurke & Scrivner, 1995; Weiss & Inwald, 2010).

Each of the different types of evaluations presents unique challenges, with regard to research, law, ethics, and practice standards. Common to all of them are a number of issues related to foundational competence. A review of the American Board of Police & Public Safety Psychology's Examination Manual shows nine foundational competencies: Professionalism, Reflective Practice/Self-Assessment/Self-Care, Scientific Knowledge and Methods, Relationships, Individual and Cultural Diversity, Ethical/Legal Standards, Practice Standards, Interdisciplinary Systems, and Knowledge of Client Milieu. Below are a number of critically important issues linked to the foundational competencies of assessment in police and public safety psychology.

Psychological evaluation of police and public safety personnel should be conducted by licensed psychologists with specialized knowledge, training, and skill (and optimally experience) that directly apply or transfer to working with law enforcement agencies and personnel (Borum, Super, & Rand, 2003; Corey, 2011; IACP, 2009; Stone, 2000). Because of the high stakes—potential harm to the officer, the agency, and the community—and the necessity to understand the complex legal and practice requirements, expertise is imperative. The case of *McCabe v. Hoberman* (1969) established that by using experts, Departments are not making decisions arbitrarily and capriciously. To that end, insofar as ABPP board certification is currently recognized by APA to be the primary recognized specialty credential in police and public safety psychology, Departments can be assured of meeting this requirement.

Police and public safety assessments should be conducted in accordance with existing guidelines (Borum, Super, & Rand, 2003). Such guidelines can be found on the International Association of Chiefs of Police website under the Police Psychological Services Section (http://theiacp.org/psych_services_section/) and in other source books in the area of forensic, I/O and clinical psychological assessment. Examiners are required to know the legal parameter s for conducting the specific types of evaluations in their jurisdiction (APA, 2002; AP-LS, 2008; Aumiller et al., 2007; Super, 1997). Finally, psychologists conducting any evaluations are to adhere to the Ethical Principles of Psychologists and Code of Conduct (APA, 2002; 2010), particularly those principles and standards related to fidelity, respecting the rights and dignity of others, and justice, but also those related to competence, confidentiality, the management of records, and the entire section devoted solely to assessment—Standard 9: Assessment. Standard 3: Human Relations also contains much in the way of guidance in terms of the foundational competencies of police and public safety psychological assessment.

Research Design and Statistics

Specialists are expected to practice in accordance with evidence-based principles. Doing so requires the ability to understand and express key elements of research design and quantitative analysis in order to evaluate research findings relative to practice in their primary domain(s) and/or to contribute to the literature through research effort. In addition, when considering the validity and reliability of psychological tests, interpreting the results of psychological tests, and/or expressing conclusions/recommendations based on those instruments and processes, specialists should be mindful of the factors that affect the accuracy and utility of their findings, as well as the evidence base supporting the use of these instruments. As scientist-practitioners, police psychologists are expected to be critical consumers of the relevant scientific literature as well as contributors to the knowledge base of our specialty through scientific endeavor.

Characteristics, principles, and limitations of common research designs and data gathering techniques

In police and public safety psychology, as in all fields of psychology, the focus of endeavor is to describe behavior, predict behavior, and to determine the causes of behavior. Different research methods are used to accomplish each of these goals. The description of behavior is the simplest goal to accomplish. Behavior can be described simply as a result of observation, although various types of descriptive data are often gathered and summarized. Descriptive strategies are often the first step in studying a particular type of behavior. However, predicting behavior, which is a bit more complex, is one of the most common goals of police and public safety psychologists operating in the assessment domain, e.g., using test, interview, and life history data to predict success as an entry-level law enforcement officer. Correlational research designs, which measure the degree of relationship between two or more variables are often used for this purpose. However, moving beyond correlation to determination of causation is the most difficult of research tasks and requires the most rigorous research designs. , In this regard, true experimental research designs are often not practical or even possible in the field of PPSP. Nevertheless, they are the "gold standard" of research, and will be discussed first before going on to other types of research design that may be used in cases where doing a classic experiment is not a realistic option.

Establishment of Causation

In order to conclude that a variable causes a particular effect, three conditions must exist (Beins, 2013). First, the variables (the presumed cause and the presumed effect) must be correlated; that is, they must co-vary in definable ways; this is referred to as the covariance rule. Although correlation between variables is necessary to establish causation, it is not sufficient. The presumed cause must precede the presumed effect;

this is referred to as the temporal precedence rule. Finally, we must rule out other variables that may have caused the effect; this is the internal validity rule. Satisfying this last criterion is the most difficult task in establishing causation, because we cannot always control for unknown or extraneous factors that might cause the effect in which we are interested. This is particularly true when conducting "in vivo" research; i.e. observing effects which occur in the real world (e.g. the police workplace) rather than in a laboratory. Nonetheless, the only type of research design that can adequately address causation is the true experimental design.

The True Experimental Design

In a true experimental design, the investigator controls and manipulates variables of interest in a systematic way, to determine whether the variables have an impact on the behavior of interest. At least two groups are created, with random assignment of participants to each group. The variable(s) that the investigator manipulates is called the independent variable(s). In the simplest case, one group experiences a manipulation (the experimental group) and one group does not (the control group or the placebo group). Then a dependent (or outcome) variable is measured in both groups to see if the manipulation produced differences in the groups. In any experiment, two hypotheses are generated to explain the results. The null hypothesis states that the manipulation of the independent variable had no effect on the dependent variable. The alternative hypothesis states that the difference in results between the groups was caused by the manipulation of the independent variable. Of course, in many experiments, the investigator believes or hopes that the null hypothesis is incorrect, and that the manipulation actually does have an effect.

To use an example, if you wanted to determine whether the amount of sleep that an officer gets affects his or her driving ability the next day, you would randomly assign officers to groups that get varying amounts of sleep (e.g., four versus eight hours), and then you would measure their ability to drive the next day to see if driving ability was affected by the amount of sleep an officer got. The null hypothesis would state that varying the amount of sleep will have no impact on driving ability. Your statistical analysis of the differences between the groups would allow you to either accept or reject the null hypothesis. Of course, you would need some way of ensuring that the officers in the different groups got the correct amount of sleep for their group assignment. You would also need to develop an appropriate and accurate way of assessing their driving ability. Additionally, you would need to carefully control any other variables that might reasonably be expected to influence driving, such as consumption of various substances. It should already be easy to see why this type of research is not often feasible in police and public safety psychology.

Correlational research

This is a form of research design that is most frequently found in the police and public safety psychology literature. This design examines the observed relationship or co-variation of two or more variables, often in the field. This design is common in psychological test validation studies, where we attempt to use test scores to predict an outcome variable. Validation coefficients typically reflect the correlations of test scores to a particular criterion, such as a performance variable. In its purest form, correlational research design is also common to the operational domain, (e.g. studies of hostage taker characteristics, studies of suicidal subjects, criminological profiling studies of known perpetrators, as they occur, coexist, or are correlated with various test scores and/or demographic descriptors).

Although relationships between the observed variables in correlational research may be measured, they are not controlled by the design. What is being measured is simply a relationship. The degree of this relationship, and the amount of the observed variance accounted for by this relationship, is measured by the statistic used (correlation, effect size, etc.). However, no cause and effect can be determined as a result of this research since any conclusions are limited to descriptions of co-variations.

This is not to say, however, that correlational research is without distinct practical value. Particularly in the operational domain, this type of correlational information is often all that is available, and it can be useful to individuals in the field who can use it to make practical/tactical decisions in real time. Correlational research can also suggest possible causes and stimulate further investigation (Mohandie, Meloy, & Collins, 2009; Mohandie & Meloy, 2010.

As noted above, there is a good deal of correlational research in the assessment domain, particularly in regard to generating validity coefficients, which are typically correlations between specific test scores, or groups of scores, with specific dependent (outcome/criterion) variables. Meta analytic research design_is a specialized class of correlational research that seeks to pull together diverse findings from several studies (usually correlational), to create broader-based theoretical evidence for various procedures or instruments (Lipsey & Wilson, 2001).

The most commonly used correlational statistics in police and public safety psychology studies include pearson product moment (r), used when measuring the relationships between two normally distributed linear variables (e.g. relationships between test scores) and Spearman (ρ, rho, or r_s), used when measuring relationships between monotonic (non-parametric) variables such as classifications or categories (e.g. differences between casualty and no-casualty outcome groups).

In addition to the *r* statistics, the results of some correlational studies may also be expressed in terms of effect size. The commonly used effect size indices are the "r" family (r^2, R^2, and eta^2) and the "d" family of effect sizes (see Cohen, 1988, Rosenthal, 1991; Rosenthal & Dimatteo, 2001). Effect size as expressed by the "d" statistic is an expression of the ratio of mean difference to pooled variability. These statistics differ from linear correlation in that they are a measure of the strength or magnitude of the relationship between two variables, (i.e. the amount of pooled variance accounted for by the relationship of the variables or means of the groups being compared). Cohen (1988) developed the following guide to interpret effect size for the d statistic when group means are being compared: < 0.1 = trivial effect, 0.1 - 0.3 = small effect, 0.3 - 0.5 = moderate effect, > 0.5 = large effect.

When evaluating correlational research results, particularly when these results are used to predict high stakes outcomes (i.e., employment suitability and/or operational competencies) as is often the case in police and public safety psychology, it is critical to evaluate both the statistical significance of the reported relationship as well as the pooled effect of the variables under study (effect size). In this regard, studies with large sample sizes and concurrent estimates of high statistical power can also yield statistically significant findings with pooled effect sizes so small as to be of little practical value (i.e., a study of relationship between marital status and work attendance with a large sample size might yield an "r" of .10 with p=.001, while this relationship would account for a very small fraction of variance, making it prone to type two error as a predictor). Conversely, too little statistical power (small sample size) can hide significant differences and effects between groups of variables (i.e., the same study with a small sample size might not recognize this correlation while this variable (marital status) if pooled with other predictors might be a useful component in a multivariate prediction equation).

Factors affecting the accuracy of conclusions of psychological research and/or predictions made by psychological tests and evaluation processes.

The validity, accuracy, and utility of conclusions, predictions, and recommendations offered by police and public safety psychologists are affected by characteristics associated with the instruments utilized, the population studied, and the criterion (dependent variable) addressed. When communicating assessment results to clients, psychologists should be sure to explain results within this context. Basic concepts, definitions, and terms relevant to this task are presented below (in summary).

Measurement error reflects the amount of an observed score or measure that is due to random extraneous variability and not part of the measurement. Classical test theory assumes that this error is random and normally distributed across all measures. As such, measurement error is the amount by which an observation or a score varies from

a true value and includes error from all sources, including sampling error inherent in test construction as well as error from other extraneous sources. One way that mmeasurement error is evaluated is though consideration of reliability and internal consistency (see Chronbach, 1951; McDonald, 1999). This is an important technical metric for evaluation of assessment instruments in Police Psychology. Although the test publisher's technical manual often contains this information in regard to the various standardization cohorts, insofar as many tests used in police pre-employment assessment are not initially constructed nor standardized on police applicant cohorts, it is also important to estimate these values when using specialized norms (i.e. police applicant norms) in these settings .

Base rate is the prevalence of an event or phenomenon within a given population. In this regard, a predictive finding derived from test scores of a given applicant may be of statistical significance from the point of view of alpha, (probability that the observed/predicted outcome is not due to chance). However, the base rate of occurrence, or beta probability (i.e., the probability that the predicted outcome will actually occur), must also be considered as a factor in determining the practical significance of this finding. In this regard, accurately predicting low base rate events (i.e., those that occur at a base rate of less than 25%) is problematic while maintaining acceptable levels of specificity (Finn 2009; McCaffrey, Palav, O'Bryant & LaBarge, 2003; Meehl & Rosen, 1955). Police applicants who successfully complete the selection process are a highly screened and range restricted population, so the base rates of most negative employment events in police psychology are usually relatively low, particularly in the first few years of employment. Therefore, when comparing a test score to a group of casualty applicants (those that experience a negative employment event), or when considering criterion-based research designs, and/or particularly when evaluating research regarding prediction of specific job outcomes, base rates should be taken into account. Failure to consider base rate of occurrence of a predicted outcome is a common cause of misclassification (see below).

Type I misclassification occurs when the null hypothesis is accepted as true but is actually false. In other words, the observed values and differences upon which the acceptance decision was based were obtained by chance. For example, based on test and interview scores a candidate for employment might be judged to not be at risk for failing training (and be screened in) yet subsequently fail training. The probability of this occurrence (Type I error) is called alpha (α) and the probability of accurate type 1 classification is (1-α).

Type II misclassification occurs when the null hypothesis is falsely rejected (i.e. the candidate is judged to be unsuitable and screened out when s/he would have been successful on the job). In other words, the observed values and differences upon which the decision was based do not represent a true difference between the means of two the classifications. For example, based on test and interview scores, a candidate

might be judged to be at risk for failing training and hence be eliminated from employment when, in fact, if hired, s/he would pass training. Base rate of occurrence is a critical factor in type 2 error. Insofar as PPSP trainees are typically selected from an applicant pool that is constricted by pre-selection factors, the base rate of failing training is typically below 20%. If the sample size upon which a study is based is also relatively small (low statistical power), the probability of type II misclassification (called β) will be high.

Managing type II misclassification is often the focus of challenges to negative employment decisions. It is a common misconception among clinicians that probability of type two errors cannot be estimated for practical reasons. In this regard, although it is usually not possible to include unselected (not hired) applicants in a straightforward outcome research design, the probability of type II error can be (and is typically) estimated through simple cross validation techniques (e.g. jackknife, bootstrap, and/or k-fold cross validation) whereby the data upon which a decision model is based is partitioned randomly into subsamples. Typically, a single subsample is retained as the validation data for testing the model, and the remaining subsamples are used as training data and to estimate sensitivity and specificity of the prediction model. (Shao & Tu, 1995).

Sensitivity and Specificity: The probability of accurate type I classification can also be expressed as the "sensitivity" of a model, and the probability of type II misclassification can also be expressed as the "specificity" of the model. The relationship of sensitivity to specificity is a measure of the efficiency of prediction.

Efficiency of prediction: As noted above, insofar as the base rate of occurrence for specific negative outcomes among public safety hires can be quite low, a model can appear to be quite accurate in regard to screening out negative outcome (sensitivity) while being quite inefficient in regard to specificity (type II error: false positive). In most prediction models used for pre-employment suitability screening, the prediction is usually expressed in terms of a binary classification (suitable/not suitable; problem/no problem; pass/fail training, etc.). Efficient models are both sensitive (screen out candidates who will have negative outcomes) and specific (do not screen out acceptable candidates by inaccurately classifying them as unsuitable). The efficiency of the prediction model is a function of the total number of accurate classifications that result as well as the number of errors that are made, particularly in regard to screening out otherwise qualified applicants (Meehl & Rosen, 1955; see also Finn, 2009, Cuttler, 2011).

Fundamental statistical techniques (descriptive, inferential, correlational, factorial)

While it is expected that all doctoral level police and public safety psychologists are familiar with the terms and concepts below, the following information is presented by way of general review. [6]

Descriptive statistics are methods of describing or summarizing data that may illustrate or suggest patterns in the data. Descriptive statistics do not allow conclusions to be made beyond the data analyzed; therefore, they cannot in themselves generate conclusions regarding hypotheses. They are simply important ways of describing the data under investigation. Descriptive statistics are often represented in frequency tables, histograms, and frequency polygons and are described in terms of central tendency and measures of spread or variability. Descriptive statistics are provided in all empirical studies, including those in police and public safety psychology. In assessment, various normative expressions (e.g. scaled scores, t scores, etc.) are derived from various descriptive statistics (central tendency, variability, etc.) in which raw scores are arithmetically converted in order to facilitate comparison across scales data sets and applicants.

Measures of central tendency include the mean, median, and mode. The arithmetic mean is the most commonly used measure of central tendency and represents the average of the values under investigation. The mean is vulnerable, however, to the effects of outliers or skewed data. In these instances, use of the median or mode may be more appropriate. The median is the middle value for a set of data that has been arranged in order of magnitude, or the value below which 50% of the scores fall. The median is less affected by outliers and skewed data than the mean. The mode represents the most frequently obtained value in the data. Measures of central tendency are presented in all empirical studies, including those in police and public safety psychology, and represent descriptive numerical summaries of the data reflected by a single "typical" number. Measures of central tendency, along with measures of spread, are basic elements of empirical study.

Common measures of spread or variability include the range, standard deviation, and variance. The standard deviation is the most commonly used statistic for describing the spread of a distribution. The standard deviation approximates the average amount by which individual values differ from the mean and is the positive square root of the

[6] An excellent review of these and other basic statistical principles may be found in Keller(2006).

variance The variance represents a measure of how far each value in the data set is from the mean. The variance is the mean of the squared deviations of the values from their mean. Measures of spread are key descriptive statistics. The standard deviation and range of scores (measures of spread) and mean (measure of central tendency) are reported in the findings of most empirical studies and typically make up the core of norms tables.

Skewness and Kurtosis may also sometimes be included as descriptive statistics. Skewness represents deviation from a normal distribution, with values skewed either to the left (low/negative) or to the right (high/positive). Kurtosis refers to the peakedness or squatness (plateau) of the distribution. Many statistical calculations assume that scores are normally distributed (follow a bell-shaped curve). Measures of skewness and kurtosis provide graphic illustration of the extent to which this assumption is correct. In some cases, scales on a personality inventory may have varying distributions (e.g. be skewed to the left or right). This can cause a T- score to have a different meaning (percentile value) across scales in which the distributions vary. This undesired result can be corrected statistically by calculation of a uniform T- score, as has been done for some of the scales on the MMPI-2.

Raw scores are converted to Z-scores to allow for the comparison of scores from different normal distributions. A Z-score indicates how many standard deviations a value is above or below the mean. *Z*-scores have a mean of 0 and a standard deviation of 1. *Z*-scores are frequently linearly converted to other types of standard scores such as T-scores, which establish a 0-99.99 theoretical range to eliminate negative scores. Since scales on personality inventories often have a different number of items per scale, raw scores by themselves do not allow for meaningful comparisons across scales. Z-scores solve this problem. T-scores are standard scores with a mean of 50 and a standard deviation of 10. *Z*-scores can be transformed into T-scores scores by multiplying the given *Z*-score by 10 (the standard deviation of the distribution of T-scores), and adding 50 (the mean of the distribution of T-scores). If the variable measured by a psychological test is normally distributed, two-thirds of the population would be expected to obtain T-scores between 40 and 60**.** T-scores are used on objective personality inventories to allow for a comparison of scores across scales. T-scores correspond to a specific population percentile with a T-score of 50 representing the 50^{th} percentile and a T-score of 60 representing approximately the 84^{th} percentile.

Norms are known population parameters on standardized tests that serve as standards of comparison for any individual who takes the test. Test publishers often present norms for a variety of groups. It is incumbent on the examiner to choose the norm group most representative of the individual being tested to avoid inaccurate interpretations. For example, the MMPI-2-RF reports norms (means and standard deviations) on a number of comparison groups including clinical, forensic, medical, and non-clinical samples. When making comparisons against a particular normative group, it

is important consider varying response set differences (e.g. pre-employment vs counseling settings) as well as the composition of the groups (applicant vs clinical cohorts). Using a dissimilar comparison group as a basis of prediction can result in serious errors. For example, making predictions from the MMPI-2-RF regarding law enforcement job performance based on clinical comparison group norms rather than personnel screening law enforcement norms can yield erroneous conclusions. Similarly, making fitness for duty determinations based on law enforcement applicant norms can fail to address the clinical nature of most FFDE referrals.

Inferential statistics are utilized when sample data are used to make inferences about populations or comparisons are being made about group characteristics.

T- tests and analysis of variance (ANOVA). These tests evaluate differences in group means. The t-test compares two group means, or a group mean to a population mean. ANOVA compares the means of multiple groups. ANOVA allows for a determination of which independent variables have a significant effect on the dependent variable and how much of the variability is attributable to each independent variable. Both methods require a single dependent variable. For example, a t-test might be used to determine if the mean scores on a specific MMPI-2 scale vary significantly under the high demand conditions of personnel selection versus low demand, low stakes conditions. ANOVA might be used to determine if the mean scores on a specific MMPI-2 scale vary according to more than two demand conditions (e.g., high vs moderate vs low vs none). ANOVA generates an omnibus test, which means that if the ANOVA is significant at least two means differ from one another. A significant ANOVA requires selection of a post hoc test (e.g., Tukey) that compares pairs of means for statistical significance.

Analysis of covariance (ANCOVA) allows for inclusion of a covariate in the analysis to account for variance that is shared between the covariate and the dependent variable. For example, a comparison of two groups of employees on job security (e.g., entry-level patrol officers vs. supervisory staff) may include age as a covariate, given that older more experienced police officers are generally more secure in their jobs than younger less experienced police officer employees, and supervisory staff tend to be older than entry-level patrol officers.

Multivariate analysis of variance (MANOVA) is an extension of ANOVA that allows for simultaneously testing the effect of the independent variables on more than one dependent variable. The application is the same as ANOVA, but is more versatile because it accommodates multiple outcome variables that are related to each other, which is the general case in research studies. For example, we may desire to determine if high scores on two specific scales on a personality inventory are associated with two independent variables of alcohol abuse (yes vs no) and absenteeism (low vs high).

Multivariate analysis of covariance (MANCOVA) is an extension of ANCOVA. Like the ANOVA-MANOVA distinction, MANCOVA is distinguished from ANCOVA by the simultaneous testing of more than one dependent (or outcome) variable. For example, expanding on the ANCOVA example described previously, we may be interested in determining if entry-level patrol officers differ from supervisory staff on measures of job security and job satisfaction, using age as a covariate.

Factor analysis is a method for identifying clusters or groups of related items, variables, and scales. Factor analysis can be used to reduce the number of variables that need to be analyzed. It also generates composites of variables (factors) that are more internally consistent, content valid, and reliable than individual variables. Exploratory factor analysis may be used in police and public safety psychology to identify the underlying constructs measured by a psychological test. Confirmatory factor analysis may be used to determine if hypotheses regarding the makeup of factors or constructs underlying a particular psychological test are observed as expected.

Multiple linear regression analysis tests for associations between a set of predictor (or independent) variables and a single, continuous outcome (or dependent/criterion) variable. Individual predictors are evaluated for statistical significance (i.e., that they explain a significant amount of variance in the outcome measure), considering the variance they share with the other predictors as well as with the outcome. Predictor variables may be included in the regression equation simultaneously (i.e., all at once), hierarchically (i.e., separately, in a pre-determined order), or statistically (i.e., predictors are entered into the equation in order of the amount of variance in the outcome they explain). The total amount of variance explained by all predictors is also evaluated for significance. For example, multiple linear regression may be used to evaluate the association of 10 separate personality scale scores collected at hiring with a scale score evaluating overall job performance at one year. Similarly, the results of multiple regression analysis may be used to identify components of classification prediction equations such as Fisher's discriminant, Linear Discriminant Analysis, Decision Tree analysis, etc. (McLachin, 2004).

Instrument Construction and Factors Affecting the Accuracy of Predictions Made by Psychological Tests and Evaluation Processes.

Assessment psychologists should be familiar with the design and technical characteristics of the assessment instruments that they use in their evaluations, as well as the assumptions, scoring rubrics, and other procedures that they use to reach conclusions and recommendations, and the limitations of each. Key topics within the following discussion include:

8. Common techniques, instruments for measurement, and/or outcome evaluation within the assessment domain.
9. Reliability (Standard error of measurement, internal consistency measures, factors affecting reliability of outcomes and measures).
10. Validity (types of validation evidence, appropriateness, and comparisons of measurement validity; building incremental validity, face validity, sources of the different types of error).
11. Instrument construction, Item analysis, and other important factors affecting utility of measurement instruments/processes (e.g. disparate impact).

Common uses, techniques, instruments for measurement, and/or outcome evaluation within the assessment domain:

The primary area of assessment in PPSP is pre-employment evaluation (PEPE) of job applicants (e.g., Kitaeff, 2011; Weis, 2010). Incumbent employees may require a fitness-for-duty evaluation (FFDE) when there is evidence that they manifest psychological conditions that could affect their ability to perform their job safely and effectively (e.g., Corey, 2011). Some agencies examine employees for special assignments (e.g., SWAT, K-9, drug task force, and undercover), and may also test employees for promotion and leadership positions. In all cases, and for all purposes, these evaluations should conform to consensually-derived best practices in terms of use of instruments, processes, and recommendations.

The IACP Police Psychological Services Section publishes practice guidelines for pre-employment assessment and fitness-for-duty evaluations (IACP, 2009). Both sets of guidelines specify the use of objectively scored psychological tests validated for use in the police population, review of relevant background and historical information, and face-to-face interview as the primary components of both (PEPE and FFDE) evaluations. In addition, whereas not all PPSP practitioners include specific measures of cognitive ability in their PEPE protocols, it should be noted that substantial scientific evidence exists to support the predictive validity of these measures (Aamodt,2004 ; Dilchert, Ones, Davis, & Rostow, 2007; Lefkowitz, 1977) in Police and Public Safety pre-employment settings. As such, many practitioners include
assessment of cognitive ability and educational achievement within their standard pre-employment protocol.

The primary domains of psychological functioning measured in pre-employment assessment via psychological tests relate to psychopathology, normal personality traits, and cognitive functioning (Weiss & Inwald, 2010; Weiss & Weiss, 2011). Consequently, the psychological tests most commonly incorporated in PPSP evaluative protocols and batteries are objectively scored tests of psychopathology (e.g. MMPI-2,

MMPI-2-RF, PAI) and personality characteristics (e.g. CPI, 16PF, NEO-PI-R). When included in the protocol, cognitive functioning is often measured by tests such as the Shipley-2 and the Wonderlic. The design of these instruments may be further delineated in terms of scale construction and criterion focus.

The scales of some psychological tests, particularly those primarily designed and standardized for use in general population settings for purposes other than pre-employment selection, often contain a number of common overlapping items. As such, the sub-scale scores on these tests are themselves correlated. Although this can be helpful in counseling and treatment planning settings, as well as in FFDE's, it has been noted that this inter correlation of personality test scales on the same instrument has the effect of inflating the apparent validity of a test and/or a prediction model in pre-employment settings, since the correlated scales are vulnerable to similar sources of error (Ben-Porath 2007; 2009; 2012; Cuttler 2011). Similarly, although correlations of scores across multiple instruments may be considered to be evidence of convergent validity, multiple predictor models and/or clinical judgments based on correlated scores across several instruments can also be expected to have higher standard error values, and, as such, will be less reliable. (Cureton, Cronbach, Meehl, et al., 1996; Campbell & Fiske, 1959).

In order to minimize internal correlational overlap, the scales on some tests (e.g. MMPI-2-RF, PAI, 16PF, NEO-PI-R) have been constructed using factor analysis in order to maximize the stability, independence of measurement and, hence, the predictive validity of these scores. Other PPSP tests (e.g., CPI 434/260) represent a hybrid approach, containing both linearly derived and factor analytically derived scores (i.e. configural analysis). Other tests are purely linear and/or have cluster derived scales (e.g., MMPI-2), which often contain common items and overlaps. Although the latter (MMPI-2) may still be useful in some PPSP assessment contexts (particularly in FFDEs) because they allow for generation of multiple diagnostic hypotheses, the limitations of these instruments should be noted when used in pre-employment assessment, particularly when making high stakes employment decisions based on multiple subscale scores on these instruments.

Similarly, many tests that were originally standardized and validated on general (nonspecific) populations were designed to be validated against psychological construct ratings. Other tests used in PPSP (e.g., IPI) were designed specifically for public safety populations, standardized on public safety populations, and validated against specific job outcomes. Both designs have their strengths and limitations, which should be considered by PPSP practitioners (Cuttler, 2011).

Several studies and meta-analytic reviews have demonstrated modest to strong correlations between scores on these tests and PPSP specific construct ratings (e.g.

Varela, Boccaccini, Scogin, Stump & Caputo, 2004; Ones, Viswesvaran, & Dilchert, 2004). It should also be noted, however, that although some of these constructs have been meticulously constructed and content linked to PPS job performance (Spilberg, 2003), these constructs are not necessarily psychometrically independent nor are they likely to be perfect predictors of actual job performance Other tests (e.g. IPI and M-Pulse) have been designed and validated to predict specific job outcomes such as lateness, auto accidents, disciplinary action, job termination, etc. However, these tests are also more narrowly standardized and, as such, scores on these tests may lead to minimized and/or exaggerated hypotheses in specific cases; i.e. small differences in applicant population scores can mask larger differences in general population scores (Ben-Porath, 2009).

PPSP practitioners might also consider the use of specialized reports and norms that are available for commonly used tests (e.g., CPI Police and Public Safety Selection Report; Roberts & Johnson, 2001). This report contains both construct-linked and job-outcome-linked predictions. Similarly, the MMPI-2-RF report available from Pearson/NCS (Ben-Porath & Tellegren, 2008) can contain both general and law enforcement specific normative scores.

Reliability (Standard error of measurement, internal consistency measures, factors effecting reliability of outcomes and measures)

Test reliability is the degree to which a test or test score consistently measures a defined construct or predicts a specific outcome (AERA, APA & NCME, 1999). At the heart of reliability measurement are the facts that test scores vary from instance to instance, and all scores obtained on a test contain both accuracy and error. It is further assumed that error scores are not correlated with true scores and that error scores vary randomly, i.e. they are equally likely to be positive or negative, so that the sum of all error scores will always be zero.

The goal of estimating reliability is to determine how much of the variability in test scores is due to error in measurement and how much is due to variability in true scores. Since we assume that errors are random and not correlated with true scores, the lower the effect of error, the higher the reliability of measurement. To estimate reliability, we take multiple scores (estimates of a construct or measures of a predictor) and observe the degree to which a given value (score) changes across these measures. If the value changes a lot we have a high degree of measurement error and a low reliability. If it does not change a lot then we assume that the effect of error is minimal (high reliability), and we can therefore attribute observed variability to differences across measurements.

Nunnally (1978) defined reliability as "the extent to which measurements are repeatable and that any random influence that tends to make measurements different from occasion to occasion is a source of measurement error" (p. 206). Although values of .70 or greater are often considered to reflect acceptable levels for test reliability, satisfactory test reliability actually varies according to the purpose of the test. Therefore, there is no absolute coefficient alpha that is considered to reflect an acceptable level of reliability. Nunnally (1978) has suggested that instruments used in basic research should have a reliability of .70 or better, while in applied settings reliabilities of .80 may be inadequate at times, and in instances involving high stakes testing coefficient alphas of at least .90 and preferably .95 are desired. Cascio (1991) similarly has suggested that in employment situations a test reliability of at least .90 is recommended.

There are several ways to estimate the reliability of a test instrument. The simplest way to measure reliability is called test-retest. However this approach is often not practical in real world testing situations such as PEPE and also does not take into account things like practice effect and changes in test-taking attitude across administrations (e.g. taking a test as an applicant and later taking it as an incumbent).

The alternate form model involves construction of two equivalent instruments that are administered to applicants who are randomly assigned to take one or the other form. This approach controls for practice effects, but is dependent on the assumption that the two tests are completely identical so that it may be assumed that observed differences in scores are results of true score and not error. In addition, the sample size must be large enough to assume equivalence between the two groups and that differences are not the effect of other extraneous variances.

The split-half model breaks a single test into two equal halves and treats each as an alternate form. The correlation between these two split halves is used in estimating the reliability of the test. This split-half reliability estimate is then stepped up to the full test length using the Spearman–Brown prediction formula (Allen and Yen, 1979). The most common internal consistency measure is Cronbach's alpha (Chronbach, 1951), which is usually interpreted as the mean of all possible split-half coefficients.

Test validity

Test validity refers to the degree to which a test measurement provides accurate extra-test information in defined domains. For a test or evaluation process to be valid, it must be demonstrated that the results generated are directly, demonstrably, and consistently related to the outcome or purpose for which the test/evaluation is utilized and that conclusions and/or recommendations that result from use of these instruments are fair, objective, and consistent with applicable employment laws.

There is no absolute standard for establishing that a test or assessment process is "valid." Rather, test/process validity is typically presented in terms of multiple sets of

"validation evidence," i.e., various measurements and experimental results derived from test data that support validation. These sets of results may be in terms of test/question item content (content validity), subject performance on specific criteria (criterion validity), independent ratings on descriptive behavioral measures (construct validity), relationships between known related measures (convergent validity), evidence of improved accuracy through combination of predictors (incremental validity), or repeated measures on independent subject groups (cross validation).

The common standard by which the quality of all validation evidence is judged is the degree to which the validation data can be shown to be both experimentally rigorous and consistent with professional standards and best practices. In this regard, SIOP (Society for Industrial and Organizational Psychology; APA Division 14) has compiled and published "Principles for validation and use of personnel selection procedures" (Fourth Edition; Society for Industrial and Organizational Psychology, 2003), which is a remarkably comprehensive professional resource document dealing with all aspects of validation as it relates to employment selection. This document presents definitions and practical subject matter discussions regarding test and process validation in great depth. The primary source documents upon which this resource is based include "Standards for Educational and Psychological Testing" (American Educational Research Association, APA, and National Council on Measurement in Education, 1999) as well as Federal employment statutes that impact the use of selection instruments in employment, particularly the Civil Rights Act Title VII (1964) and related EEOC guidelines[7].

All Specialists in police and public safety psychology practicing within the assessment domain should be thoroughly familiar with the contents of these documents on a detailed level, and all applicants for Board Certification in PPSP are encouraged to review these documents.

The degree to which a particular test or pre-employment assessment process is considered "valid" is a function of both the amount and quality of the validation evidence presented. In the pre-employment setting, virtually all validation measures (e.g., content, construct, criterion, incremental, convergent, etc.) should be referenced to the characteristics and critical performance attributes of the job in question. The common vehicle for describing these job attributes is the job analysis. According to Muchinsky (2012), a thorough job analysis documents the tasks that are performed on the job, the situation in which the work is performed, and the human attributes needed to perform the work. Human attributes, as such, are often defined as "knowledge skills and abilities (KSAs) necessary to perform the job in question.

[7] See EEOC - Uniform Guidelines On Employee Selection Procedures (1978)

There are many ways to develop and analyze job descriptive information. Robinson (2012) notes that the three most popular job analysis methods are the Critical Incident Technique (CIT), Hierarchical Task Analysis (HTA), and the Position Analysis Questionnaire (PAQ). CIT is the least structured approach, in which person-oriented data is derived through interviews with incumbents and other subject matter experts (SMEs) who are asked to give examples of positive and negative performance incidents. In HTA, a job is analytically divided into a series of functions, tasks, and subtasks, often through a focused process of on the job observation. The PAQ approach is a structured approach based on data generated by the Position Analysis Questionnaire (McCormick, Jeanneret, & Mechem, 1972). Although it is labeled a questionnaire, the PAQ is actually designed to be completed by a trained job analyst who interviews the SMEs (e.g., job incumbents and their supervisors).

There are a number of other approaches to job analysis, as well. Functional job analysis (Fine & Cronshaw, 1999) is a classic example of a task-oriented technique in which work elements are scored by panels of SMEs in terms of relatedness to data (0–6), people (0–8), and things (0–6). However, in all cases, the result is a detailed description of job performance parameters as well as the KSAs necessary to perform the job. When addressing the question of validity, particularly in regard to pre-employment assessment, PPSP practitioners should be prepared to describe the manner in which their findings and predictions relate to the job as described in analytical terms.

It is incumbent upon PPSP practitioners to be able to critically evaluate the published validation evidence regarding the instruments and processes they employ, such as information found in publisher's technical manuals, as well as in other published research. Additionally, inasmuch as test manuals and other research that present validation evidence on general (community) populations can be expected to result in different validities in PPS specific populations, cross validation comparisons are critical. Similarly, tests and processes that are validated on PPSP populations nationally can have different characteristics than local populations, and hence differential validities in local settings. As such, in addition to consideration of published technical information, all PPSP specialists should gather local data and cross validate findings.

Finally, it is important to take into account how base occurrence rates of criterion variables (e.g. personality problems, demonstrated psychopathology, and/or specific negative job outcomes) and other sources of measurement error affect test validity (e.g., Finn, 2009). Similarly, range restriction of test scores, predictor variables, and criteria may affect a measure's validity (Hunter & Schmidt, 1990). Furthermore, inherent limitations related to measured sensitivity, specificity, and positive predictive power are important to consider. As noted in an earlier section , in order to produce the most valid predictions from the test data, psychologists are generally encouraged to utilize empirical or statistical methods (Cuttler, 2011; Grove et. al., 2000; Meehl, 1954),

but they may need to incorporate clinical judgment when necessary (Westen & Weinberger, 2004).

Other important factors affecting utility of measurement instruments/processes (e.g. disparate impact)

In addition to being valid and reliable, all tests and selection procedures used in employment settings must be shown to be free of disparate or adverse impact against specific gender and/or racial groups. According to US Civil Rights Act Title VII and the EEOC Uniform Guidelines (1978), disparate (or adverse) impact is defined as follows: "A selection rate for any race, sex, or ethnic group which is less than four-fifths or 80 percent of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact"[8].

Although the 4/5 (a.k.a 80%) rule stated above is the most common way of calculating adverse impact, since it is simple and easy to calculate, this simple calculation may overestimate the existence of adverse impact in a practical sense, once again due to the differential base rates associated with "minority" vs "majority" groups (Roth, Bobko & Switzer, 2006). In this regard, Biddle (2005) describes a number of other statistical tests (e.g. variants of mean difference Z-scores expressed as chi square functions) that compare differences between means and/or other applicant pool distribution characteristics across minority and majority groups. However, Biddle also notes that although these tests may minimize type I error (i.e., the error of determining adverse impact exists even when it does not), these measures are not particularly accurate in regard to type II error (indicating adverse impact does not exist when, in fact, it does).

Once again, and as discussed in previous sections, this failure to account for type II error in measuring adverse impact is a function of the relatively low level of statistical power in most applicant groups (i.e., the lower base rate of occurrence of females and minorities in the applicant pool). In this regard the typical incidence of women and racial minority groups reported by Law Enforcement agencies is usually below 15% and often quite a bit lower. Although statistical measures of adverse impact may be useful in determining the existence of adverse impact in large groups (more than 2500 applicants), they may overestimate the existence of adverse impact in smaller applicant pools. As such, in a practical sense PPSP psychologists practicing in the pre-employment assessment domain are encouraged to evaluate the tests and processes they choose to use in terms of the largest groups possible. Often this may initially involve reliance on test publisher's technical reports. However, it is also important that practitioners in this area systematically aggregate their own data across time, and

[8]EEOC – Uniform Guidelines On Employee Selection Procedures (1978), § 1607.4 para D

develop methods and protocols for evaluating adverse impact in real time (see Morris 2001)

Assessment Domain, Reference List[9]

Aamodt, M. G. (2004). *Research in law enforcement selection.* Boca Raton, FL: Brown Walker Publishing.

AERA, APA & NCME (1999). *Standards for educational and psychological testing*. Washington, DC: AERA, APA, and NCME.

Allen, M.; Yen W. (1979). Introduction to Measurement Theory. Monterey, CA: Brooks/Cole.

Americans with Disabilities Act of 1990, Pub. L. No. 101-336, §2, 104 Stat. 328 (1991).

American Psychological Association (2002). Ethical Principles of Psychologists and Code of Conduct. American Psychologist, 57, 1060-1073.

American Psychological Association (2013). Specialty Guidelines for Forensic Psychology. American Psychologist, 68, 7-19.

Arkes, H. R.; Hammond, K. R. 1986. Judgment and decision making: An interdisciplinary reader. Cambridge University Press, New York, NY, US.

Aumiller G., Corey, D., Allen S., Brewster J., Cuttler, M., Gupton, H., & Honig, A. (2007). Defining the field of police psychology: Core domains & proficiencies. *Journal of Police & Criminal Psychology*, 22, 65–76.

Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, NJ: Prentice Hall.

Beck, A. T., Ward, C.H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Arch. Gen. Psychiatry*, 4 (6), 561–71.

[9] References cited herein are considered to be representative of the core science subject matter discussed. This list is NOT exhaustive and should by no means be considered a comprehensive reading list for specialty board examinations or competence.

Beck, A., Weissman, A., Lester, D., & Trexler, L. (1974). The measurement of pessimism: The Hopelessness Scale, *Journal of Consulting and Clinical Psychology*, 42, 861–865.

Beins, B. C. (2013). *Research methods: A tool for life* (3rd ed.). NY: Pearson.

Ben-Porath, Y. S. (2007, October). *Use of the MMPI-2 Restructured Form in assessing law enforcement candidates.* Paper presented at the meeting of the Police Psychological Services Section of the International Association of Chiefs of Police. New Orleans, LA.

Ben-Porath, Y.S. (2009). The MMPI-2 and the MMPI-2-RF. Unpublished manuscript, California Peace Officer Standards and Training Commission, Sacramento, California.

Ben-Porath, Y.S. (2012). *Interpreting the MMPI-2-RF*. Minneapolis, MN: University of Minnesota Press.

Ben-Porath, Y.S,. & Tellegen, A. (2008). T*he Minnesota Multiphasic Personality Inventory -2 Restructured Form: Manual for administration, scoring, and interpretation*. Minneapolis, MN: University of Minnesota Press.

Biddle, D. (2005). *Adverse Impact and Test Validation: A practitioner's Guide to Valid and Defensible Employment Testing*. London England: Gower.

Binet, A. & Simon, T. The Development of Intelligence in Children: The Binet-Simon Scale. Baltimore, MD, Williams and Wilkens Company.

Blau, T. (1994). Psychological Services for Law Enforcement. New York, John Wiley and Sons, Inc.

Borum, R., Super, J., & Rand, M. (2003). Forensic assessment for high-risk occupations. In A. M. Goldstein & I. B. Weiner (Eds.), Forensic Psychology (Vol. 11 pp. 133-147). Hoboken, NJ: John Wiley.

Burgess, E.W. (1928). Factors determining success or failure on parole. In A.A. Bruce (Ed.), The workings of the indeterminate sentence law and the parole system in Illinois. Springfield: Illinois State Board of Parole.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81-105.

Cascio, W. F. (1991). *Applied psychology in personnel management* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Civil Rights Act of 1964, Title VII, Pub. L. 88-352 (78 Stat. 241).

Civil Rights Act of 1991, Pub. L. 102-166.

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

Corey, D.M. (2011) Principles of Fitness-for-Duty Evaluations for Police Psychologists, in J. Kitaeff, (Ed.) *Handbook of Police Psychology* (pp 263-293). New York, NY: Routledge.

Corey, D. M. (2012). Core Legal Knowledge in Police & Public Safety Psychology. Paper presented at the American Board of Professional Psychology Summer Workshop Series, Boston, MA, July 11, 2012.

Corey, D., & Borum, R. (2013). Forensic assessment for high risk occupations. In R. K. Otto (Ed.), Forensic psychology (pp. 246-270). Vol. 11 in I. B. Weiner (Editor-in-Chief). Handbook of psychology (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Corey, D. M., MacAlpine, D. S., Rand, R., Rand, D. C., & Wolf, G. D. (1997). *B-PAD for Oral Boards: A Guidebook for the Entry-level Police Version* (2nd ed.). Napa, CA: The B-PAD Group, Inc.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334.

Cureton, E. E., Cronbach, L. J., Meehl, P. E., Ebel, R. L., & Ward, A. W. (1996). Validity. In A. W. Ward, H. W. Stoker, & M. Murray-Ward, (Eds.), *Educational measurement: Origins, theories, and explications, Vol. 1: Basic concepts and theories* (pp. 125-243). Lanham, MD: University Press of America.

Cuttler, M. J. (2011). Preemployment screening of police officers: Integrating actuarial prediction models with practice. In J. Kitaeff, *Handbook of Police Psychology (pp. 135-163).* New York, NY: Routledge.

Dawes, R.M. (1988). Rational choice in an uncertain world. , Harcourt Brace Jovanovich, San Diego, CA, US.

Dawes, R.M., Faust, D., & Meehl, P.E. (1989). Clinical versus actuarial judgment. Science, 243,
1668–1774.

Dilchert, S., Ones, D., Davis, R.D., & Rostow, C.D. (2007). Cognitive ability predicts objectively measured counterproductive work behaviors. *Journal of Applied Psychology*, 92 (3), 616-627.

Enforcement guidance: Pre-employment disability-related inquiries and medical examinations under the Americans With Disabilities Act of 1990. (1995). Equal Employment Opportunity Commission, ADA Division, Office of Legal Counsel. Washington, DC.

Faust, D., (1984) The Limits of Scientific Reasoning. University of Minnesota Press. Minneapolis, MN, US.

Faust, D. & Ahern, D. S. (2012). Clinical judgment and prediction. In D. Faust (Ed), Coping with Psychiatric and Psychological Testimony (pp. 147-208) (6th ed.) . Oxford: Oxford University Press.

Faust, D., Ziskin, M (2012). Coping with Psychiatric and Psychological Testimony (6th Edition) Oxford: Oxford University Press.Fine, S. A. & Cronshaw, S. F. (1999). *Functional job analysis: A foundation for human resources management*. Mahwah, NJ: Erlbaum.

Finn, S. E. (2009). Incorporating base rate information in daily clinical decision making. In J. N. Butcher (Ed.), *Oxford handbook of personality assessment* (pp. 140-149). New York, NY, Oxford University Press.

French, John R. P. (1953). Experiments in Field Settings. In L. Festinger & D. Katz (Eds.), *Research Methods in the Behavioral Sciences*. Chicago ILL Dryden Press.

Grove, W.M., & Meehl, P.E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. Psychology, Public Policy, and Law, 2, 293–323.

Grove, W.M., Zald, D.H., Lebow, B.S., Snitz, B.E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, 12, 19-30.

Genetic Information Nondiscrimination Act of 2008 (GINA). Pub.L. 110-233, 122 Stat.881, 42 U.S.C. 2000.

Goldberg, L. R. (1970) Man versus model of man: a rationale plus evidence for a method of improving on clinical inferences. Psychological Bulletin, 73, 422-432.

Goldberg, L. R. (1992). The development of markers for the big-five factor structure. Psychological Assessment, 4, 26-42.

Goldberg, L. R. (1993). "The structure of phenotypic personality traits". *American Psychologist* **48** (1): 26–34. doi:10.1037/0003-066X.48.1.26. PMID 8427480.

Grove, W.M., & Meehl, P.E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. Psychology, Public Policy, and Law, 2, 293–323.

Hanson, R. K., & Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessment for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological Assessment, 21, 1-21.*

Hathaway, S. R., & McKinley, J. C. (1940). A multiphasic personality schedule (Minnesota): I. Construction of the schedule. Journal of Psychology, 10, 249-254.

"Hawthorne Experiments." Encyclopedia of Business. Ed. Jane A. Malonis. Vol. 1. Gale Cengage, 2000. eNotes.com. 5 May, 2013, http://www.enotes.com/hawthorne-experiments-reference/

Heilbrun, K; DeMatteo, D; Marczyk, G; Goldstein, A M. (2008) Standards of practice and care in forensic mental health assessment: Legal, professional, and principles-based consideration. Psychology, Public Policy, and Law, Vol 14(1), Feb 2008, 1-26. doi: 10.1037/1076-8971.14.1.1

Hogarth, R. M. (1987). Judgment and choice (2nd ed.). Wiley, New York, NY, US.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

International Association of Chiefs of Police. (2009). Psychological Fitness-for-Duty Evaluation Guidelines. Arlington, VA: International Association of Chiefs of Police.

Kahneman, D., Slovic, P., & Tversky, A. (1982). Judgment Under Uncertainty: Heuristics and Biases. New York: Cambridge University Press.

Kates, S. (1950). Rorschach Responses, strong blank scales and job satisfaction among police officers. *Journal of Applied Psychology*, 34 (4), 249-254.

Kitaeff, J. (Ed.). (2011). *Handbook of Police Psychology*. New York, NY: Routledge.

Keller, D. K. (2006). *The tao of statistics: A path to understanding (with no math).* Thousand Oaks, CA: Sage.

Klopfer, B., (1946) The Rorschach Technique: A Manual for a Projective Method of Personality Diagnosis, World Book Co, (Yonkers-on-Hudson).

Koch, S. (Ed.). (1959). Psychology A Study of Science. Vol.3. New York, NY: McGraw-Hill Book Company, Inc.

Kuhn, T. S. 1962. The structure of scientific revolutions. Chicago: University of Chicago Press.

Kurke, M & Scrivner, E. (1995). Police Psychology into the 21st Century. Lawrence Erlbaum Associates New York

Lefkowitz, J. (1977). Industrial-organizational psychology and the police. *American Psychologist*, 32(5), 346-364.

Leonel v. American Airlines, 400 F.3d 702 (9th Cir. 2005). Retrieved from http://caselaw.findlaw.com/us-9th-circuit/1224462.html

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis. Applied social research methods series (Vol. 49)*. Thousand Oaks, CA: Sage Publications.

Mahoney, M. J. (1974). *Cognition and behavior modification*. Oxford, England: Ballinger.

McCaffrey, R. J., Palav, A. A., O'Bryant, S. E., & LaBarge, A. S. (2003). *Practitioner's guide to symptom base rates in clinical neuropsychology*. New York, NY: Kluver Academic/Plenum.

McCormick, E. J., Jeanneret, P. R. & Mecham, R. C. (1972). A study of job characteristics and job dimensions as based on the Position Analysis Questionnaire (PAQ). *Journal of Applied Psychology*, 56(4), 347-368.

McCrae, R.R., & Costa, P.T. (1987) Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, 52, 81-90.

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

McLachlan, G. J. (2004). Discriminant Analysis and Statistical Pattern Recognition. Wiley Interscience. ISBN 0-471-69115-1. MR 1190469.

Meehl, P.E. (1954). *Clinical vs. statistical prediction*. Minneapolis, MN: University of Minnesota.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, 52, 194–216.

Meichenbaum, D. (1977). *Cognitive behavioral modification: An integrative approach.* New York, NY: Plenum Press.

Mohandie, K. & Meloy, J. R. (2010) Hostage and barricade incidents within an officer-involved shooting sample: Suicide by cop, intervention efficacy, and descriptive characteristics. *Journal of Police Crisis* Negotiations, 10: 1, 101 — 115

Mohandie, K., Meloy, J. R., & Collins, P. (2009). Suicide by cop among officer involved shooting cases. *Journal of Forensic Sciences*, *54*, 1–7.

Morris, S. B. (2001). Sample size required for adverse impact analysis. *Applied HRM Research, 6*, 13-32.

Muchinsky, P. M. (2012). *Psychology Applied to Work*. Summerfield, NC: Hypergraphic Press, Inc.

Nisbett, R., & Ross, L. (1980). Human inference: Strategies and shortcomings of social judgment. Prentice Hall, Englewood Cliffs, NJ, US.

National Advisory Commission on Criminal Justice Standards and Goals. (1973). Retrieved from http://onlinebooks.library.upenn.edu/webbin/book/lookupname?key=United%20States.%20National%20Advisory%20Commission%20on%20Criminal%20Justice%20Standards%20and%20Goals.

Novaco, R.W. (1975). Anger control: The development of an experimental treatment. Lexington, KY: Lexington.

Nunnally, J. C. (1967). *Psychometric theory* (1st ed.). New York, NY: McGraw-Hill.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Ones, D. S. & Viswesvaran, C. (2003). Job-specific applicant pools and national norms for personality scales: Implications for range-restriction corrections in validation research. *Journal of Applied Psychology, 88,* 570-577.

Ones, D.S., Viswesvaran, C. & Dilchert, S (2004, November). A construct-based, comprehensive meta-analysis and implications for pre-offer screening and psychological evaluations. Paper presented at the meeting of the International Association of Chiefs of Police (IACP) Los Angeles, CA.

Pre-Employment Psychological Evaluation Guidelines. (2009) Police Psychological Services Section of the International Association of Chiefs of Police.

Principles for validation and use of personnel selection procedures (4th Ed.). (2003). Society for Industrial and Organizational Psychology. Washington, DC: American Psychological Association.

Psychological Fitness-for-Duty Evaluation Guidelines. (2009). Police Psychological Services Section of the International Association of Chiefs of Police.

Reese, J. (1995). A history of police psychology services. In Kurke, M. & Scrivner, E. (Eds.) Police psychology into the 21st century. Hillsdale, NJ, Lawrence Erlbaum Associates. 31-44.

Report of the National Advisory Commission on Civil Disorders. (1967). Retrieved from http://faculty.washington.edu/qtaylor/documents_us/Kerner%20Report.htm

Roberts, M. & Johnson, M. (2001). *CPI Police And Public Safety Selection Report Technical Manual* (1st ed.),Los Gatos Ca Law Enforcement Psychological Services, Inc.

Records of the Enforcement Assistance Administration (LEAA) (1965-1977). Retrieved from http://www.archives.gov/research/guide-fed-records/groups/423.html

Robinson, M. (2012). What is Job Analysis? Institute of Work Psychology. Retrieved from. http://esrccoi.group.shef.ac.uk/pdf/whatis/job_analysis.pdf

Roth, P. L., Bobko, P. L., & Switzer, F. S. III. (2006). Modeling the behavior of the 4/5ths rule for determining adverse impact: Reasons for caution. *Journal of Applied Psychology, 91,* 507-522.

Rosenthal, R. (1991). *Meta-analytic Procedures for Social Research*. Newbury Park, CA : Sage.

Rosenthal, R., & DiMatteo, M.R. (2001) Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Rev. Psychol*., 52, 59-82.

Sawyer, J. (1966). Measurement and prediction, clinical and statistical. Psychological Bulletin, 66, 178-200.

Shao, J & Tu,D (1995). The Jackknife and Bootstrap. Springer-Verlag, Inc. pp. 281

Society for Industrial Organizational Psychology, Inc. (2003). Principles for the validation and use of personnel selection procedures, fourth edition. Bowling Green, OH: Author.

Spilberg, S.W. (2003, April 11). Development of psychological screening guidelines for police officers: Background and development of essential traits. In S. W. Spilberg & D. S. Ones (Chairs), Personality work behaviors of police officers. Symposium conducted at the 18th annual meeting of the Society for Industrial and Organizational Psychology, Orlando, FL.

Spielberger, C.D., Gorssuch, R.L., Lushene, P.R., Vagg, P.R., & Jacobs, G.A. (1983). *Manual for the State-Trait Anxiety Inventory*. Palo Alto Ca Consulting Psychologists Press, Inc.

Stern, G.G., Stein, M.I., and Bloom, B.S. (1956). Methods in personality assessment. Free Press, Glencoe, IL.

Stone, A.V. (2000). Fitness for duty: Principles, methods and legal issues. Boca Raton, FL: CRC Press.

Super, J. (1997). Select legal and ethical aspects of fitness for duty evaluations. Journal of Criminal Justice, 25, 223-229.

The challenge of crime in a free society: A report by the president's commission on law enforcement and administration of justice. (1967). United States Government Printing Office, Washington, D.C. Retrieved from https://www.ncjrs.gov/pdffiles1/nij/42.pdf

The Civil Rights Act of 1991. Retrieved from http://www.eeoc.gov/laws/statutes/cra-1991.cfm

The Omnibus Crime Control and Safe Streets Act of 1968. Retrieved from http://transition.fcc.gov/Bureaus/OSEC/library/legislative_histories/1615.pdf

Tupes, E.C., & Christal, R.E. (1961) Recurrent Personality Factors Based on Trait Ratings. Technical Report ASD-TR-61-97, Lackland Air Force Base, TX: Personnel Laboratory, Air Force Systems Command.

Uniform guidelines on employee selection procedures. (1978). Retrieved from http://www.gpo.gov/fdsys/pkg/CFR-2011-title29-vol4/xml/CFR-2011-title29-vol4-part1607.xml

"Uniform guidelines on employee selection procedures". uniformguidelines.com. Retrieved November 14, 2007.

Varela, J.G., Boccaccini, M.T., Scogin, F., Stump, J., & Caputo, A. (2004). Personality testing in law enforcement employment settings: A metaanalytic review. *Criminal Justice and Behavior,* 31, 649 -675

Vrieze, S. I., & Grove, W. M. (2009). Survey on the use of clinical and mechanical prediction methods in clinical psychology. *Professional Psychology: Research and Practice*, 40(5), 525-531.

Weiss, P. A. (Ed.). (2010). *Personality assessment in police psychology: A 21st century perspective*. Springfield, IL: Charles C. Thomas.

Weiss, P.A. & Inwald, R. (2010). A brief history of personality assessment in police psychology. In P.A. Weiss (Ed.) *Personality Assessment in Police Psychology: A 21st Century Perspective*. 5-28 Springfield, IL: Charles C. Thomas.

Weiss, P.A., & Weiss, W.U. (2011). Criterion-related validity in police psychological evaluations. In J. Kitaeff (Ed.), *Handbook of police psychology*. 125-133 New York, NY: Routledge.

Westen, D. & Weinberger, J. (2004). When clinical description becomes statistical prediction. *American Psychologist*, 59, 595-613.

Wiggins, J.S. (1973) Personality and prediction: Principles of personality assessment. Addision-Wesley Publishing Company, In. Philippines.

Wiggins, J.S. (2003). *Paradigms of Personality Assessment.* New York, NY: Guilford.