**A Bayesian Approach to Item Development**

**Working with Alternative Item Types**

Over the past few decades, forensic psychologists who employ self-report clinical personality measures have seen a significant increase in impression management by examinees who are well aware of the purpose of the testing and gear their responses to serve their agenda. For example, it is not uncommon to find huge elevations on scales measuring obsessive-compulsive personality traits among individuals who area actually impulse-ridden, emotionally volatile, and psychologically disorganized.

If such subjects are being assessed for custody of a child or for criminal sentencing purposes, their agenda is to present themselves as behaviorally overcontrolled, inhibited, accommodating, and respectful of rules and laws. High social desirability scales such as Obsessive-Compulsive invite dishonest examinees to register large elevations. This type of impression management, however, is often picked up by the test’s internal validity scales, thus partially or completely invalidating the test record.

With no alternative test procedures that are resistant to examinee impression management, evaluators working with this type of client are deprived of psychometric findings to bolster conclusions developed on the basis of interview impressions and case history data. During the 1950s and 1960s when self-report clinical testing methods began to enjoy enormous popularity among psychologists and were the subjects of a very large number of research studies, a rival form of assessment, projective techniques, also rapidly ascended in popularity among clinicians. During the heyday of classical psychoanalysis, it was not uncommon for analysts to refer patients for assessment with projective techniques in order to provide the analyst with insight on problematic aspects of the treatment.

It was also around this time that psychologists were called upon more and more often to render opinions on psychopathology in legal matters, whereas this function had formerly been the exclusive province of medical practitioners. The buttressing of the psychologist’s testimony by test results, including both self-report clinical personality testing and the results of projective techniques, appears to have contributed to the legitimization of psychologists as expert witnesses in criminal, civil, and family court.

Although clinical psychologists have for many decades regarded projective techniques as being highly informative and valid means of clinical assessment, empirical research studies have yielded either negative or mixed results with respect to the validity of these techniques. It should be noted that there is a seemingly impassable divide between the way projective techniques are employed by many clinicians and the demands of a truly valid empirical research basis for these instruments. Although “blind” interpretation of projective test protocols (where the evaluator does not have access to details of the case) is sometimes used in didactic settings, the majority of projective test experts strongly prefer to interpret findings in conjunction with specifics of the clinical situation. Thus, in addition to numerical scores that have been developed, especially with various Rorschach systems, clinicians regard content analysis of subjects’ responses as providing avenues to a deep understanding of aspects of the psyche that the subjects may not even be fully aware of themselves.

While previous reviews of projective drawings methods research concede that they have some degree of validity as a general index of psychological disorganization or distress, those authors do not regard drawings as offering much help in the diagnosis of specific disorders (Swensen, 1957; Kahill, 1984; Lilienfeld, Wood & Garb, 2000).

Part of the difficulty in prior efforts at turning drawing methods data into a viable assessment instrument is the lack of an appropriate psychometric model for this type of data. Classical test theory (CTT) (Nunnally, 1967) and item response theory (IRT) (Suen, 1990 ) are best suited for sets of items that are administered together as a test, or in the case of IRT, individual items developed under the assumption of an underlying unidimensional trait. In the case of CTT, the most common method of assigning scores to test performance is to locate examinees’ raw scores a certain number of standard deviation units from the mean of test scores in the standardization sample. In IRT, the subject’s ability is determined by performance on a succession of individual items calibrated according to parallel scales of subject ability and item difficulty.

In the case of mental ability and academic achievement testing, where IRT is now the dominant paradigm, scores are assigned according to subject’s performance on sets of items with known characteristics. Where computerized adaptive testing is the procedure, subjects are assessed as having a particular standing on the trait of interest based on their performance on a succession of items calibrated according to the particular model used, whether one, two, or three parameter or a more recent complex variation. Probabilistic methods are employed in determining the point at which subjects’ abilities are estimated within a sufficiently narrow band specified by the needs of the testing program.

With respect to self-report clinical personality testing, examinee standing on particular traits is expressed in the form of T-scores, locating the examinee’s performance a certain number of standard deviation units away from the sample mean, as noted above. The one exception to this format is the Millon Inventories, where examinees’ scores are directly linked to the external criterion of clinician diagnoses. That is to say that in the Millon system, the BR (base rate) score includes cut points at which clinicians begin to diagnose subjects as having traits or features of particular personality disorders and at a higher point where clinicians begin to diagnose subjects as having the full-blown personality disorder.

Millon named the BR score after concepts articulated by Meehl and Rosen (1955) in a seminal article that stressed the importance of considering base rates in making clinical diagnoses, and presents scale validities in terms of classification efficiency statistics that include Sensitivity (SENS), Specificity (SPEC), Positive Predictive Power (PPP), and Negative Predictive Power (NPP), not merely correlation coefficients. The Millon Inventories manuals, however, do not include explicit discussions of changing the probability that the subject has the target condition based on test performance (posterior probability).

As useful and informative as projective techniques may be, the dependence of this type of interpretation upon specifics of the individual case renders these approaches incapable of being evaluated empirically. While this may be less of a problem with respect to purely clinical assessments, at the present time it constitutes a major obstacle to the use of projective techniques in forensic psychology. A major setback to the use of such techniques in court settings occurred in 1993 with the United States Supreme Court decision *Daubert versus Merrill-Dow Pharmaceuticals, Inc.* (1993). According to the *Daubert* decision, novel scientific techniques must pass a four-prong test in order to be admitted as scientific evidence in federal courts.

The first prong imposed by the *Daubert* decision requires that the novel procedure must enjoy general acceptance in the relevant scientific community (which would appear to conflict with the designation of the procedure as “novel”). The procedure must also be the subject of an article or articles in refereed scientific journals. The procedure must have a known error rate. Finally, the experiments purporting to demonstrate the scientific acceptability of the procedure must be designed in such a way as to be “falsifiable.” This is a somewhat confusing term of art in the philosophy of science (see Popper, 1959) that in simple terms means the experimental design can prove the hypothesis false as well as confirming it. An example of a falsifiable experimental design might be a study of the hypothesis that individuals who have been diagnosed with Schizophrenia, Paranoid Type are more likely to engage in assaultive behavior than are individuals with no psychiatric diagnosis.

The experimental procedure might consist of recording police contacts or institutional staff reports on individuals with a diagnosis of Schizophrenia, Paranoid Type and comparing the frequency of assaultive incidents in that group with similar records of a group of individuals who do not have that diagnosis. If the difference in frequency of assaults between the two groups is above a certain statistical threshold, then we may say that the null hypothesis is rejected and a difference does exist in the population from which these two samples are drawn. If, however, there is only a negligible difference, then we may state that the proposition is false, if there is sufficient statistical power in the experimental design. We can even “accept” the null hypothesis at *beta* (defined as 1-power).

An example of an experimental proposition that is possible to falsify might be a test of an observation of Sigmund Freud’s that concerning the aggressive character of the superego. Freud held that the superego has an aggressive, punitive character because in the process of internalization of the parental object that forms the behavioral model for the superego, the original tie to the parental love object is replaced by an identification. This results in the withdrawal of a quantum of libido, leaving a residue of aggression. The inherent difficulties of devising an experimental procedure to prove that proposition false are self-evident. Similarly, the intertwined numerical scores, response content, case details, and theoretical propositions that characterize projective test interpretation place these techniques in the category of being hopelessly unable to be falsified.

What lay jurors want to know with respect to test findings is the likelihood that the individual in question has the psychological condition that is a factor in the legal matter at hand. This question appears to be addressed by the PPP statistic, which gives the percentage probability that if the test says that the subject has the condition, the individual actually does have that condition. Another classification efficiency statistic that would seem to offer a satisfactory answer is SENS, which gives the probability that an individual who has the target condition will also have a positive score on the test. In other words, these two classification efficiency statistics respectively answer the questions 1) how likely is it that if the test says the person has it, the test is right, and 2) if an individual has the condition, how likely is it that the test will pick it up?

Relying on either of these two statistics alone can be extremely misleading, however. As persuasive as high SENS may be, those findings can be entirely negated by other characteristics of the sample. For example, if there many subjects scoring positive on the predictor variable, then despite very high SENS (most cases are picked up by the predictor) there may also be so many false positives that SPEC is poor, resulting in only a minimal increase in the overall probability of the condition, given the predictor. This necessity of taking into account baseline frequencies, not merely relying on SENS to the exclusion of SPEC, is discussed in detail in Meehl & Rosen.

Many experienced assessment experts are unfamiliar with the nuances of operating characteristics. Two such experts sharply criticized Theodore Millon when the MCMI-II was published and NPP and SPEC for the various scales were included in the test manual. Their criticism was that the specificities in the high nineties for many of the scales were impossibly high and must have been misreported or were the result of error. They did not consider the fact that with instruments that have a wide variety of clinical scales, most standardization subjects do not have the target condition measured by any individual scale and most subjects do not have a clinically elevated score on that particular scale.

To offer an extreme example, if we standardized a test in a sample that included only Antisocial Personality Disorder and Generalized Anxiety Disorder subjects and the test consisted solely of psychometric scales measuring these two conditions, we would not expect those who had Anxiety Disorder to present with the recklessness and impulsivity of Antisocial Personality Disorder, as these subjects are typically fearful and inhibited. Neither would we expect the antisocial subjects to experience much anxiety. Based on these expectations, both the Anxiety and Antisocial scales should have very high SPEC, given the characteristics of the sample.

A classical view of psychometrics would be biased against any predictor that had extremely poor SENS or extremely poor SPEC. After all, the standard for assessing Sensitivity or Specificity should be chance, and a chance procedure such as a coin flip would produce a finding of 50% SENS and 50% SPEC. However, even if a predictor has only modest SENS, it can still yield a very high classification accuracy if SPEC is very high and it correctly classified the great majority of the sample who do not have the condition.

The present article describes a method for working with “projective “ test data from a strictly empirical perspective, excluding any influence from clinical theory and any reliance on specific case details to shape interpretations. The procedure was designed to meet the requirements of *Daubert*, including publication in a peer reviewed journal, known error rate, and falsifiability. The theoretical underpinning for GAVEL is Bayes’ Theorem, which in terms of psychometric research, may be described as the mathematically optimal procedure for determining the conditional probability of a criterion variable given a predictor variable.

In Bayesian terms, the baseline probability of the criterion is its prior probability, and the conditional probability, given a positive score on a predictor variable, is its posterior probability. If additional predictor variables are employed with an individual criterion variable, the posterior probability reflecting the change in probability associated with the first predictor variable may be used as then new prior probability (baseline) for the next predictor variable, with the process repeated for additional predictors. Bayes’ Theorem includes SENS, SPEC, probability of the predictor variable, and baseline probability of the criterion in calculating posterior probability.

Westbury (2010) notes that clinical psychologists have largely ignored the advice of Meehl & Rosen to pay attention to base rate scores as a factor in using psychometric tests to make diagnostic and other decisions. Psychometric theorists, on the other hand, have employed Bayesian methods in a variety of situations, including IRT models (Suen, 1990; Price, 2017). To the author’s knowledge, neither clinical psychologists nor psychometric experts in the field of test design have employed Bayes’ theorem in its most basic form, that is the computation of posterior probability based on new information, in the design of any sort of psychometric instruments.

In contrast to IRT, which requires a unidimensional construct with highly correlated items for optimum model fit, the present method requires the opposite. Bayesian predictors that are even moderately correlated with one another are likely to produce inflated posterior probability estimates when used sequentially.

In the early days of projective techniques, examiners would make absolute interpretations of the sort “this means this.” In other words, their model was strictly binary, offering a specific interpretation when the characteristic appeared in a protocol, without taking into account any possibility that the test subject did not actually have those characteristics even though they registered the sign on the test. This style thrived because projective testing was born of the authoritarian psychoanalytic tradition in which expert opinion was respected and not questioned. Echoes of this approach are to be seen even today in clinical and forensic reports where projectives are used.

Therefore, even though projective techniques have the advantage of being less susceptible to conscious distortion because characteristics of a pathological response are not obvious in the way that they are in self-report questions where social desirability may be readily inferred from item content, the lack of empirical validity and absolutist interpretation style represent huge obstacles in forensic work. This is especially true since *Daubert* and its decades of influencing state as well as federal courts.

The GAVEL method of treating psychometric data from drawings techniques was designed to overcome the problem of clinical test invalidity because of impression management through recasting a very old “projective” technique on a strictly empirical basis that has nothing to do with projection. Steps in the development of GAVEL are discussed in detail in Dyer (2018). Briefly, these include selection of initial development of initial item pool from the standard clinical repertory of diagnostic signs from the House Tree Person literature and additional material from the author’s own experience with the procedure, correlating all items in the initial pool with a variety of clinical and psychometric criteria, selecting items significantly correlated with these criteria, and making editorial adjustments. The resulting reduced item pool now totaling 66 predictors was then cross validated in a study with a new sample.

Additional research was conducted to compare the performance of GAVEL with clinical scales of the Personality Assessment Inventory in the prediction of behavioral and diagnostic criteria (Dyer, submitted for publication). All of the above research was conducted with unit weights for each of the GAVEL predictors. Taking into account the fact that the predictors included in GAVEL vary widely in the resulting degree of change in the probability that an individual belongs to a particular diagnostic or behavioral group, the final test development phase consisted of Bayesian analyses of all 66 predictor items against relevant subsets of 15 behavioral and diagnostic criteria, depending on the particular item’s usefulness in the prediction of each criterion variable.

This analysis involved the computation of a posterior probability for the criterion, given the individual predictor item. The combined sample in which this final phase of the test development project was conducted included 497 subjects who had been given the House-Tree-Person scored by the GAVEL method and whose case records were reviewed for the presence or absence of each criterion variable. While some predictors had a reasonably high frequency and most of the criterion variables also had a reasonably high frequency, there remain many instances in which the frequencies of the predictor or number of subjects scoring positive on particular criterion variables were not large enough to provide a sufficiently stable estimate of the exact change in probability of the criterion variable. Therefore, rather than expressing the outcome as a posterior probability figure, GAVEL presents results in the form of a **posterior probability change score** **in percentage intervals of seven points.** This change score, termed a B7 score, represents one score point for each one-to-seven point increase in posterior probability of the criterion.

GAVEL therefore rests on a psychometric foundation that is the most basic form of application of Bayesian methods. It seeks to capture the change in the likelihood that an individual has a particular condition or behavior, given an aggregation of probabilities associated with specific predictors of that condition that appear in the GAVEL record

Most early attempts at validating House-Tree-Persion indicators involved the study of individual items. This placed the research at a severe disadvantage, as individual predictors have limited validity in comparison to psychometric scales composed of multiple items. While some of this early research did employ scales (made up of items selected on the basis of content and theoretical expectations instead of empirical methods), the individual scale elements were given unit weights. This resulted in the loss of valuable information, as drawing indicators have widely differing levels of correlation with external criteria.

GAVEL addresses these problems by combining B7 scores for each criterion variable. This provides the advantages of 1) multiple rather than single predictors for a particular criterion variable, and 2) taking into account the degree to which individual predictors change the baseline probability. This use of B7 scores in aggregate is based theoretically on Bayesian sequential posterior probability calculation, using each posterior probability as the new prior probability for the next predictor variable.

GAVEL offers interpretation of specific predictor-criterion relationships expressed in terms of likelihoods: mildly suggestive, strongly suggestive, strongly indicative, etc. The interpretive inference model is that individuals who have an above-baseline probability of membership in a particular criterion group such as assault, anxiety, depression, or schizophrenic spectrum disorder tend to display characteristics associated with that group. For example, subjects with high B7 scores for Anxiety may be described as prone to nervousness and inhibition. Those with high B7 scores for Greater than Three Arrests may be described as likely to have contempt for rules and laws, poor impulse control, and deficient capacity for anticipating consequences of actions.

There is a further procedure for aggregating total B7 scores for individual criteria based on similarities among sets of individual criteria. Summary scores are grouped under Severe Mental Illness, Substance Problems, Distressing Symptoms, Criminal Tendencies, Physical Aggression, and Child Maltreatment. Summary aggregations are based on content, and have not yet been studied empirically

The system described above is intended to provide an appropriate means of dealing with data from procedures that do not have set groups of items that are presented to subject, but predictors that appear spontaneously and naturally as the procedure is administered. In regard to GAVEL specifically, his is because original projective HTP is unstructured the only cues that the subject receives are being told what to draw. Thus, IRT and CTT psychometric models do not easily apply to such data. It is the author’s view that the present system would also be appropriate for other “projective” instruments such as the Rorschach, the validity of which has perhaps been even more fiercely criticized than that of drawings methods.

In summary, the data analytic system is probabilistic, with some similarities to the model underlying IRT. One important difference, however, is that in the Bayesian GAVEL system, it is necessary for predictors NOT to be more than weakly correlated, opposite of IRT and CTT. GAVEL employs the most basic application of Bayes’ Theorem, the calculation of posterior probability. B7 scores are completely empirical and NOT based on theory or subjective interpretation of content. Items were developed solely on basis of a conservative statistical approach to determine empirical relationships. This is true even in cases where the relationship is counterintuitive because there is no link to a theoretical relationship between drawing sign and criterion. Current GAVEL criteria are static, derived from the inspection of subjects’ case records, with no psychometric instruments used as criteria in the test-test type of validity research so common in self-report clinical personality measures research.

References

*Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993)

Dyer, F.J. (2017) Prediction of psychiatric hospitalization, diagnoses, arrests, and violent behavior through scored drawings and associations. *Psychological Reports*, *121*, 1, 4-25.

Kahill, S. (1984). Human figure drawings in adults: An update of the empirical evidence. *Canadian Psychology, 25*, 269-292.

Lilienthal, S.O., Wood, J.M. & Garb, H.N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest, 1*, 27-66.

Meehl, P. & Rosen, A. (1955). Antecedent probability and the efficiency of diagnostic signs, patterns, or cutting scores. *Psychological Bulletin, 52,* 3, 194-216.

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

Popper, K. (1959). *The logic of scientific discovery*. London: Routledge.

Price, L.R. (2017). *Psychometric methods: Theory into practice*. New York: Guilford Publications.

Suen, H. K. (1990). *Principles of test theories.* Hillsdale, NJ: Lawrence Erlbaum and Associates.

Swensen, C.H. (1957). Empirical evaluations of human figure drawings. *Psychological Bulletin, 54*, 431-465.

Westbury, C.F. (Nov. 16, 2010). Bayes’ rule for clinicians: An introduction. *Frontiers of Psychology* (online journal).