To appear in Handbook of Individual Differences
Tomas Chamorro-Premuzic, Adrian Furnham and Sophie von Stumm (Eds.)
This version is the final, as submitted version and
will differ from the published version.

Methodological Advances in Differential Psychology

William Revelle
David M. Condon
Joshua Wilt
Northwestern University

Methods for differential psychologists are the methods of all scientists: describe and test models of data. Our field is distinguished by the nature of our data and the specialized tools we use for analysis. For the differential psychologist, data come from self-report, from observations, from physiology, and from behavioral residues. Data are recorded over time and over space. The challenges of collecting data are limited only by our imagination. Methods of analysis emphasize model fitting and model evaluation.

The goals of methods in Differential Psychology are no different from those of any other science: descriptive and testable explanations of phenomena. Methods thus involve the collection and analysis of data. What distinguishes scientific fields from each other, and the field of differential psychology in particular is what constitutes data, the theories of our data, and the analytical techniques used to describe and model data. This chapter is divided into two main sections: the kinds and sources of data we collect and the ways in which we model (analyze) the data. In that entire text books are devoted to data collection, to design (Shadish et al., 2001), to inference (Pearl, 2000), and to each of many ways to model data (Judd et al., 2009; Loehlin, 2004; McArdle, 2009; McDonald, 1999; Mulaik, 2010; Rasch, 1960), this review will necessarily be of the basic concepts rather than the specifics of particular methods. For a thorough discussion of research methods of individual differences that is limited to personality narrowly defined (e.g, not including intelligence, interests or values) see the handbook edited by Robins et al. (2007).

contact: William Revelle revelle@northwestern.edu
Draft version of December 20, 2010
This is the precopy edited draft. The final version will differ slightly.

$$Data = Model + Error \tag{1}$$

A revolution in data analysis has occurred over the past thirty years: the recognition that we *model* data and compare alternative models to each other (Rodgers, 2010). This approach is summarized in Equation 1 which, if we recognize that our error is someone else's signal, is better expressed as Equation 2:

$$Data = Model + Residual. \tag{2}$$

The process of research then is one of finding models that fit the data with acceptably small residual values. "Models, of course, are never true, but fortunately it is only necessary that they be useful. For this it is usually needful that they not be grossly wrong." (Box, 1979, p 2). The current approach goes beyond just asking for usefulness by asking if the specified model is better than alternative models (Rodgers, 2010).

Coomb's Theory of Data and Cattell's Data Box

The left hand sides of Equations 1 and 2 are Data. What are the data that we collect? At an abstract level, data can be organized along three different dimensions: type of comparison (order versus proximity), the elements being compared (people, objects, people x objects) and the number of comparisons (one or more) (Coombs, 1964). Within this framework, a person can be said to be more than an object (e.g., if passing an ability test item) or

to be near an object (if endorsing an attitude item), and one person can prefer one object to another object (be closer to one attitude than another) or have a stronger preference than someone else. People can also differ in the way they group objects. The Coombs (1964) model continues to be used within psychometrics by virtue of the distinction between ability and preference items in item response theory (Chernyshenko et al., 2007) and in terms of individual differences in multidimensional scaling of situational stress.

Cattell's *data box* (Cattell, 1946) emphasized three sources of data: People, Tests, and Occasions and considered how correlations can be taken between tests and across people at one occasion (R analysis), just as correlations can be found between people across tests (Q analysis), or tests can be correlated within people across occasions (P analysis), etc. Subsequently, Cattell (1966) expanded the data box to include Background or preceding variables as well as Observers. The data box concept has been used throughout differential psychology to demonstrate the many ways of analyzing data, but the primary influence has probably been on those who study personality and cognitive development and change over the life span (McArdle & Bell, 2000; Mroczek, 2007; Nesselroade, 1984).

Methods of data collection

Individual differences can be assessed by asking people about themselves (their identity) and other people (their reputation) or by observing behavior (what people or other animals do), physiology and behavioral residues. Of these, the predominant method is probably that of self report, through the use of either questionnaires, projective instruments, or narratives.

Self report

"Do you get angry easily?", "Do you find it difficult to approach others?", "Do you make people feel at ease?", "Do you do things according to a plan?", "Do you carry the conversation to a higher level?". These are typical self report items taken from the International Personality Item Pool (IPIP, Goldberg, 1999). They follow the basic principle that if you want to know something about someone, ask them. With the instruction to answer the way you normally behave, these measures of trait Neuroticism, Extraversion, Agreeableness, Conscientiousness, and Openness show stability over long periods of time and correlate with suitable behavioral observations and other reports (Roberts et al., 2007). In contrast to measures of ability, these items are thought to measure typical performance. In other words, they measure how one usually thinks, feels and behaves rather than how *well* one can think.

A similar example would include self-report items that allow inference about the internal states of Energetic Arousal or Tense Arousal (Schimmack & Reisenzein, 2002; Thayer, 2000). When asked about energetic arousal (how alert, active or vigorous one feels in contrast to sleepy, tired or drowsy) or tense arousal (anxious, worried or tense versus calm or relaxed), subjects' scores will change over the day and in response to factors such as caffeine, exciting or depressing movies, and exercise (Revelle, 1993).

These items are direct and obvious. They may be formed into scales using factorially homogenous keying (Goldberg, 1972), also known as an inductive strategy (Burisch, 1984). Classic examples of such inventories are the Eysenck Personality Inventory (EPI, Eysenck & Eysenck, 1968), the NEO-PIR (Costa & McCrae, 1985), and the sixteen Personality Factors (16PF, Cattell & Stice, 1957). Some inventories, however are developed using the empirical or external strategy of finding items that distinguish known groups from people in general, e.g., the MMPI (Hathaway & McKinley, 1943) or the Strong Vocational Interest Inventory (Strong, 1927). They also differ from rational or deductively constructed tests such as the California Psychological Inventory (CPI, Gough, 1957) or the Personality Research Form (PRF, Jackson, 1967).

The advantages and disadvantages of empirical, rational, and homogenous keying techniques were well reviewed by Goldberg (1972), Hase & Goldberg (1967) and Burisch (1984). In general, rational and factorial techniques work better for predicting more predictable criteria, but empirical/external techniques are better able to predict very unpredictable criteria (e.g., dropping out of college). Tests assessing interests (Holland, 1959, 1996; Strong, 1927) have traditionally used empirical scale construction methods and have incremental validity when predicting diverse criteria such as success in graduate school (Kelly & Fiske, 1950).

Some question how self reports can be valid given the tendency to dissimulate or self enhance. R. Hogan & Nicholson (1988), R. Hogan & Kaiser (2005), and R. Hogan (2007) address this issue for predicting real life criteria (leadership effectiveness in organizations). Self report measures are quite successful at predicting this important criterion. J. Hogan et al. (2007) directly address the problem of faking and report that it was not a problem for selecting job applicants for security positions.

Constructing self report inventories. Practical advice for constructing self report inventories for the differential psychologist (e.g., Clark & Watson, 1995; Simms & Watson, 2007; Watson, 2005) emphasizes starting with a good theoretical understanding of the constructs to be measured and the population of interest, writing items that are clear and readable, examining the internal structure of the items, purifying the scales developed, checking for external validity in terms of correlations with criterion groups, further refinement of items and finally extensive documentation. Issues to consider include breadth of items, definition of facets of the construct, clarity of wording of items, response analysis using IRT technique, suitability for the target population and evidence for convergent, discriminant, and construct validity. Types of item selection techniques include empirical based upon known groups, homogeneous, based upon the factor/cluster structure of the domain of items, or just rational choice based upon theory.

Narratives

Narrative approaches to individual differences have grown in popularity in recent years. Researchers collecting narrative data typically do so as a means to assess how people make sense out of their lives (Pasupathi & Hoyt, 2009). Therefore, the preferred units of analysis are life-stories or discrete scenes from one's life-story. Many narrative researchers work from the perspective of *narrative identity* (McAdams, 2008): from this perspective, the psychological construction and telling of a life story brings together one's remembered past and imagined future into a narrative identity that potentially provides life with some degree of unity, meaning, and purpose (Singer, 2004). Life stories feature particular scenes occurring at different times in one's life and, like any good story, convey a variety of themes through its structure, characters, and plot (McAdams, 1993).

Due to the massive amount of scenes, events, and memories a person accumulates throughout a lifetime, quantitative analysis of narrative identity at first seems a daunting undertaking. Indeed, the cumbersome methods of the case study and the study of single lives are more amenable to qualitative analysis. However, modern narrative researchers have been up to the task, as the past two decades have seen steady growth in creative, quantitative methodologies to analyze narratives.

One fruitful approach to dealing with the problem of scene selection is the introduction of the standardized life story interview (McAdams et al., 1997) in which people narrate a set of important scenes in their lives (high points, low points, turning points, vivid memories from childhood, adolescence, adulthood, and an imagined future scene) and trained human coders assess these scenes for structural and thematic elements. Studies employing this approach aggregate scores for such themes as emotional tone, complexity, and coherence (McAdams, 1993). Another approach for analyzing narratives, which focuses on the importance of individual scenes rather than the entire story, is to have people narrate a self-defining memory (Singer & Blagov, 2004). Self-defining memories are especially emotional and vivid scenes that communicate how people came to be who they are today and may be coded similarly to the scenes in the life story interview. An innovative method of assessing narrative data is to code how people think about their own narratives, termed autobiographical reasoning (Habermas & Bluck, 2000). The process of autobiographical reasoning is analogous to telling a meta-narrative, as people reflect and comment on the meaning of different scenes in their own narratives and what implications those scenes may have (McLean, 2005). Still others obviate the need for human coders by taking advantage of the ability of computerized text analysis programs to count words relevant to various thematic categories (Pennebaker et al., 1997). For example, researchers interested in how much positive emotional content is conveyed in a narrative have the ability to count how many positive emotion words such as happy, joy, or elated appear in their participants' narratives.

Ability tests

The typical self report inventory measures what people normally do. Ability tests measure how well they can do. Originally developed as predictors of poor school performance, ability tests such as the SAT and GRE have become standard predictors of college and graduate student performance (Kuncel et al., 2001, 2007). Commercial IQ tests are given in most clinical assessments. Within the field of cognitive abilities, there have been two broad traditions, the psychometric measurement oriented approach and the cognitive processes approach. With a better understanding of the cognitive processes involved in ability tests, it is thought possible to combine cognitive theory with advanced psychometric principles (e.g., Item Response Theory) to create more efficient testing instruments (Embretson, 1998). Unlike the open source IPIP (Goldberg, 1999) there does not seem to be a public domain set of ability items that different labs can use. Rather, there are sets of commercial tests, both individualized and group forms that need to be purchased, or "home brew" tests that are unique to particular lab groups.

A fundamental assumption of ability tests is that performance is not affected by motivational state and that all participants are performing at the best of their ability. This is, however, not true. See Revelle (1993) for compelling evidence that motivational states associated with caffeine or diurnally variable energetic arousal affects ability test performance by up to one standard deviation. Individual differences in anxiety and stereotype threat have also been shown to affect cognitive performance, even on high stakes testing.

Other report

The ratings of professional psychologists (Fiske, 1949), of teachers (Digman, 1963), of peers (Norman, 1963, 1969; Tupes & Christal, 1961), or of self show a remarkable degree of consistency in identifying 5 broad factors of behavior (Digman, 1990). These five have become known as the 'Big 5' dimensions of personality (Digman, 1990; Goldberg, 1990). However, not all find such a simple five dimensional solution. Walker (1967) when comparing teacher, peer and self ratings among elementary school children showed consistency in identifying a two dimensional circumplex structure with primary axes that could be interpreted as activity and neuroticism. With the use of appropriate internet techniques, it is relatively easy is to get useful informant reports (Vazire, 2006).

Behavioral observation

Self-report and to a lesser extent other-report have been the most prominent ways of assessing personality, however, perhaps the most intuitive way to do so is to observe how people actually behave. This sound reasoning underlies the use of behavioral observation. Although intuitive, behavioral observation has rarely been employed, due in part to the relatively high costs associated with devising a viable behavioral observation scheme (Funder, 2001). Indeed, it is much more difficult to develop a system for coding behavior, train coders, and actually conduct observations than it is to have individuals or informants fill out global personality ratings (Furr & Funder, 2007). Notwithstanding these costs, behavioral observation is worth pursuing for the simple reason that actual behavior is what psychologists really care about (Baumeister et al., 2007). Thus, behavioral observation may be held as a gold standard in differential psychology.

Behavioral observation may occur in natural settings or in laboratory settings. A longstanding goal of differential psychology is to predict what people do in naturally occuring environments, however, it is obviously difficult to collect such data in a non-intrusive way. A new methodology called EAR (Mehl & Pennebaker, 2003) relies on a small recording device that is programmed to turn on and off throughout the day, recording for a few minutes at a time, producing objective data in natural environments. Laboratory based methods of behavioral observation by definition lack some of the external validity of naturalistic studies but offer controlled environments in which to examine behavior. The German Observational Study of Adult Twins (GOSAT) project of Borkenau et al. (2001) has had participants take part individually in structured laboratory activities designed to elicit behaviors relevant to the Big 5. Extending Borkenau et al. (2001) methods, Noftle & Fleeson (2010) have recently reported the first results of a large scale observational study of people interacting in group activities; these studies observed not only content of behavior but how much behavior varies as a function of age across adulthood. Behavioral observation in the lab is not limited to adults, as exemplary studies

conducted by Emily Durbin and colleagues (Durbin et al., 2007; Durbin & Klein, 2006; Durbin et al., 2005) have used standard laboratory tasks designed specifically to elicit behavior related to childhood temperamental characteristics.

In each of the aforementioned studies, researchers had to make difficult decisions about what to observe. Indeed, no one study is large enough to catalogue all behaviors; thus, it is important to carefully consider theoretical reasons for choosing variables. Oservational studies may assess discrete behaviors (e.g., smiles) by counting the frequencies of their occurrence, or by having observers make a single rating of a target on behavioral dimensions (Borkenau et al., 2004). Coding systems for behavior/emotion are available, with the Riverside Behavioral Q-Sort (Funder et al., 2000) and the Facial Action Coding System (FACS) developed by Ekman et al. (1978) as perhaps the best known and well-validated measures. Choices also have to be made about how many observers to employ, who should observe the target behavior, and whether observation should be done live or from videorecordings (Furr & Funder, 2007). These choices should be guided by the theoretical questions each study is attempting to answer. It is also important to assess the quality of coded data; indices of inter-rater agreement are typically computed as intraclass correlations (Shrout & Fleiss, 1979), which may be computed in various ways in order to best suit the structure of one's coding system. The recent increase in commitment to behavioral observation and advances in technology making this method more feasible are moving differential psychology toward a becoming a more mature science of actual behavior.

Physiological measures

The utilization of physiological measures is typically done with the purpose of discovering the biological basis or etiology of individual differences (Harmon-Jones & Beer, 2009). Neuroimaging techniques are among the most popular physiological measures employed; the specific neuroimaging technique used in a particular study depends on the theoretical question the study is designed to investigate. Researchers interested in how brain *structure* relates to individual differences rely on Magnetic Resonance Imaging (MRI) in order to generate detailed images of the brain (DeYoung et al., in press). Studies concerned with brain *activity* may use functional MRI (fMRI) (Canli, 2004). fMRI relies on the blood oxygen level dependent (BOLD) contrast effect to measure blood flow as an indicator of brain activity. Another way that differential psychologists measure brain activity (D. L. Johnson et al., 1999) is Positron Emission Tomography which detects gamma rays emitted from a tracer introduced to the body to generate images. fMRI and PET have good spatial resolution but poor temporal resolution; therefore, researchers interested in measuring brain processes as they occur (Wacker et al., 2006) may prefer to use electroencephalography (EEG). EEG records electrical activity along the scalp generated by neurons firing in brain and has good temporal resolution but poor spatial resolution. A popular physiological measure outside of the brain is salivary cortisol (Chida & Steptoe, 2009), which relates to Hypothalamic Pituitary Axis stress-response. Other physiological measures showing reliable individual differences include body temperature (Baehr et al., 2000), blood pressure, heart-rate, skin conductance, and eye-blink startle response (Diamond & Otter-Henderson, 2007).

Remote data collection

Perhaps the most challenging methodological question for personality researchers is the desire to assess individual differences in a manner that holistically reflects all the relevant aspects of personality through the use of assessment tools with fine-grain accuracy. In fact, this is generally not possible due to limitations regarding the number of items that individual participants are willing to take. The historical resolution of this challenge has been the pursuit of accurate data which is limited to a unique domain. Today, it is possible to meet this challenge through the use of remote data collection procedures and the combination of responses from vastly greater sample sizes. The technique of *Synthetic Aperture Personality Assessment* (Revelle et al., 2010) gives each participant a small subset of items from a larger item pool, and then combines these responses across subjects to synthetically form very large covariance matrices.

The main source of remote data collection comes from survey-oriented, web-based studies. Though the use of internet samples is appealing in terms of the ease of collection and the diversity of samples (Gosling et al., 2004), this relatively new method

does present some unique challenges. Of considerable importance is the implementation of safeguards against the incidence of repeated participation by the same subject. The incidence of more insidious concerns (such as misrepresentation or item-skipping) is more difficult to avoid and must therefore be taken into account during data analysis (J. A. Johnson, 2005). In addition, traditional paper-and-pencil measures do not always transfer to electronic formats without distortion and, even when such migrations are possible, care must be taken to maintain validity (Buchanan et al., 2005). To this end, a large number of scales are accessible in the public domain through the central IPIP repository (Goldberg et al., 2006).

While the web-based studies are the primary source of growth within the use of remote data collection, several other technologies contribute to this methodology. Some of these measures are addressed below in the context of longitudinal studies. Notably, recent advances in "self-tracking" technologies provide more reliable replacements to diary-based studies of behavioral and affective measures. One example of this technology is the electronically activated recorder (EAR) employed by Mehl et al. (2007). Research based on the use of this device to date have explored differences in the conversational habits across gender and well-being.

National and international surveys

One consideration for researchers who are interested in exploring individual differences in longitudinal research is that data from some studies are openly accessible. For instance, the U.S. Bureau of Labor Statistics allows free access to the results of several longitudinal surveys (though some datasets may require application). Examples of these studies include the National Longitudinal Survey of Youth (NLSY79), which has tracked about 13,000 young men and women since 1979 and their biological children since 1988 (Harvey, 1999). Many other countries (including Britain, Australia, Korea, Switzerland, Canada and Germany) offer comparable datasets which are openly available or can be accessed through the Cross-National Equivalent File (Burkhauser & Lillard, 2007). Of course, many research topics are not amenable to the use of pre-existing datasets. When appropriate however, these resources can be a practical and invaluable means of conducting longitudinal or cross sectional analyses in a fraction of the time that is typically required.

In addition to these longitudinal data sets, large scale assessments often make use of multiple data collection methods. The Programme for International Student Assessment (PISA) for example, employs both survey methods (for collecting information about participants' backgrounds and opinions) and behavioral methods (for testing participants' aptitude in mathematics, reading, science and problem-solving skills). The data from PISA assessments, which are conducted with 15 year old participants every three years in as many as 65 countries, are disseminated by the OECD and freely available for analysis (Anderson et al., 2007). See (e.g., Hunt & Wittmann, 2008) for an examination of the relationships between national intelligence, levels of educational attainment and national prosperity. A variety of other topics are covered through similar assessments by national and international agencies, including the International Monetary Fund, the World Health Organization and the United Nations. Despite lacking the flexibility of customized designs, use of such data allows for insightful comparative analyses across countries and large groups.

Animal research

As it has with other fields, the study of animal behavior offers individual difference researchers the opportunity to design experiments which would be impractical or unethical with human subjects (Vazire et al., 2007). Until recently the use of animal research to study differential psychology was primarily in lesion and drug studies (e.g., Gray, 1982; Gray & McNaughton, 2000) or in multigeneration selection studies for reactivity in the rat (Broadhurst, 1975). Observational studies of ongoing behavior in non-human animals in unrestricted environments has been relatively limited, having been constrained by measurement challenges (Gosling & Vazire, 2002) and the "specter of anthropomorphism" (Gosling & John, 1999). Research to date has included such obvious subjects as dogs and chimpanzees in addition to more surprising choices, such as snakes and octopuses (Gosling, 2001) or the pumpkin seed sunfish (Coleman & Wilson, 1998). Such animal research are currently limited to the use of observational behavioral reports and include a number of unique challenges (Vazire

et al., 2007). It is likely however that the ability of animal research to contribute to the study of human personality will increase over time as best practices are identified and further developed.

Types of designs

As has been ruefully commented upon many times (Cronbach, 1957; Eysenck, 1966; Vale & Vale, 1969), the broad field of psychology has represented two seemingly antithetical approaches: the *experimental* and the *observational*. Reconciliations and unifications of these approaches have been repeatedly called for (Cronbach, 1957, 1975; Eysenck, 1997) with limited success (Revelle & Oehleberg, 2008). Both approaches have the same goal: to identify (causal) sources of variance unconfounded with other variables.

The classic difference between these two approaches has been an emphasis upon central tendencies versus variation, between statistics emphasizing group differences (t and F) versus those emphasizing variation and covariation (σ^2 and r). But with the realization that these statistics are all special cases of the *general linear model* it became clear that the difference was not one of analysis, but rather of theory testing.

Experimental approaches

The essence of an experimental approach is *random assignment* to condition. Randomization serves to break the correlation between experimentally manipulated *Independent Variables* (IVs) from non-observed but potentially *Confounding Variables* (CVs). The set of potentially confounding variables is infinite, but includes individual differences in age, sex, social status, education, prior experience, and motivation as well as situational variables such as time of day, immediate past experience, interactions between subject variables and experimenter characteristics (e.g., sex of subject interaction with sex of experimenter). By randomly assigning participants to experimental conditions, the expected value of the correlation of the IV with the CVs is zero. Although never actually zero, as sample size increases, the unobserved confounding correlations will tend towards zero.

Person by condition interactions

Experimental approaches to the study of individual differences would seem oxymoronic, for how can we randomly assign individual differences? We can not. But we can investigate the relationship between individual differences and the experimentally manipulated conditions to test theories of individual differences. The power of interactions between individual differences (sometimes called *Person Variables* or PVs) and our experimental IVs is that the PV * IV interaction allows for a clearer understanding of the limits of the effects of both. Interactions show the limit of an effect. By having an interaction, we can rule out many extraneous explanations. That introversion is associated with better performance on exams could be because introverts are smarter than their more extraverted colleagues. But with a stress manipulation that reverses the rank orders of introversion and performance, we can rule out an ability explanation (Revelle et al., 1976).

Between-person vs. Within-Person. Individual differences researchers study factors that vary across individuals (between-person variability) and factors that vary across time and situation within the same individual (within-person variability)[1]. It is important to realize that, although the between-person relationship for two variables will mirror the within-person relationship for those variables in some instances, this is not a necessarily the case (Fleeson et al., 2002). Thus, for the same reason that questions pertaining to between-group and within-group relationships must be analyzed separately, so must investigations of between-person and within-person relationships.

Lab based

The power of interactions of a experimental variable with an individual difference variable was shown in a series of experimental studies examining the effect of caffeine induced arousal on cognitive performance. Rather than finding any main effects of individual differences or of caffeine it became clear that caffeine enhanced performance for some of the people, some of the time. The first study

[1] Sometimes between-person variability is referred to as interindividual variability, whereas within-person variability is referred to as intraindividual variability

in this series showed that caffeine and time pressure hindered the performance on a complex test similar to the Graduate Record Exam about .6 standard deviations for the most introverted participants while simultaneously enhancing performance about the same amount for the more extraverted participants (Revelle et al., 1976). This was initially taken as evidence in favor of the arousal model of extravesion (Eysenck, 1967). But with further examination, this effect was true only in the morning, and only true for the impulsivity sub-component of extraversion (Revelle et al., 1980). This led to a rethinking of the arousal model as well as to a reconceptualization of the measurement of extraversion (Revelle & Revelle, 1981). Indeed, further experiments involving the interactions of anxiety with feedback manipulations, and the demonstration of the independence of these effects from the caffeine effects led to a theory integrating trait and state individual differences with situational stressors and cognitive processes (Humphreys & Revelle, 1984).

Lab based studies have long been a staple of research investigating Reinforcement Sensitivity Theory (**??**). Recent studies attempting to integrate theories of functional impulsivity with RST (Smillie & Jackson, 2006) and test whether fear and anxiety originate from separable neurobehavioral systems described by RST (Perkins et al., 2007) continue in this tradition. Additionally, research on individual differences in anxiety (Wilt et al., in press) exemplify the wide range of experimental methods available (Armstrong & Olatunji, 2009; Fox et al., 2001) to differential psychologists.

Randomized field studies

Although typically associated with lab based studies, experimental design also enhances field studies (Cook et al., 1979). Consider the effect of anxiety on student performance in a gateway science course (in this case, a year long course in biology is a requirement for a major in biological sciences, Born et al., 2002). Prior work had suggested that performance is enhanced for women and minority students when assigned to study groups. To avoid confounding with a 'volunteer effect', Born et al. (2002) examined how study groups interacted with anxiety and gender by randomly assigning volunteers to study groups or a control condition. At the end of the year they were able to disentangle the study group effect (by comparing those randomly assigned to study groups and their randomly matched controls) from the volunteer effect (by comparing volunteers not assigned to study groups with non-volunteers).

Many long term health studies have randomly assigned participants to condition. When analyzing these data, it is tempting just to include those who follow the research protocol. Unfortunately, this is where individual differences become very important, for it has been found that conscientious placebo takers have reduced mortality rates compared to their non-adherent counterparts (Gallagher et al., 1993; Horwitz et al., 1990; Irvine et al., 1999) That is, the behavioral correlates of personality can swamp any effects due to an experimental manipulation.

Observational approaches

In contrast to experimental studies which can examine the causal relationship between environmental manipulations and individual performance, observational studies try to infer latent states based upon the covariance of various measures at one time or the patterning of results across time.

Cross sectional studies

Far more than any other type of design, cross-sectional studies represent the predominant approach for researching individual differences. When employed to its full potential, a single cross-sectional design has the power to capture a wide variety of correlations across multiple domains and emphasize the relevance of individual differences in the process. Most of the published literature reflects this approach and does not need to be discussed here.

Longitudinal studies

Though substantially outnumbered by cross-sectional designs, longitudinal studies have played a crucial role in the evolution of differential psychology as a field. The primary reason relatively few researchers have employed longitudinal designs historically is because they require a greater commitment of resources and are therefore thought to introduce incremental risk, especially in academic environments where funding is uncertain and career development is often tied to publication. However,

it's also the case that carefully constructed longitudinal studies can be considerably more powerful than cross-sectional designs and that this incremental power should be taken into account when comparing the merits of both approaches (Costa & McCrae, 1992). While longitudinal studies may introduce new confounding variables, they typically reduce the variance of cross-sectional measures of a given construct by virtue of repeated measures. More importantly, they allow researchers to gather data on many topics (e.g. the stability of traits over the lifespan) which cannot be adequately addressed with cross-sectional approaches.

Longitudinal methods represent "the long way" of studying personality (Block, 1993), and in some cases those lengths have extended well beyond fifty years. Though able to inform a number of important issues, the explicit - and perhaps most important - goal of these long-term studies is to identify the factors that lead to longer and healthier lives. For instance, several prominent examples of longitudinal research have explored the relationship between intelligence, morbidity and mortality, a field recently referred to as cognitive epidemiology (Deary, 2009).

Based on the Scottish Mental Health Surveys of 1932 and 1947 and subsequent follow-ups, findings from Deary et al. (2004) demonstrate the higher intelligence levels in youth are predictive of both survival and functional independence in old age. An earlier example is Terman's Life-Cycle Study, which began in 1921 and tracked high IQ schoolchildren until their deaths (Friedman et al., 1995; Terman & Oden, 1947). Though measures used by Terman were less developed than those in use today, they were progressive for their time and sufficient for correlating life expectancy outcomes with subsequently developed personality constructs such as the Big Five. Most notably, the findings include correlations between longevity, conscientiousness and a lack of impulsivity.

Within the field of cognitive epidemiology, many researchers are using longitudinal methods to further specify the factors which mediate life outcomes. In terms of the differential effects of maturational and generational changes, Elder (1998) has performed comparative analyses across longitudinal studies with the Terman Life-Cycle Study and the Berkeley Institute studies, which tracked children born approximately 10 and 20 years after the "Termites" (Block, 1971; Elder, 1998). On the basis of the age differences across these samples, Elder has focused his analysis on the differential developmental impacts of the Great Depression and World War II (Elder, 1998; Elder et al., 1994). In the case of WWII, sample participants who were older when entering military service paid a higher price in terms of health outcomes and career interruption than those who entered at younger ages (Elder et al., 1994). His findings suggest that even global, historical events of this nature can have non-uniforms effects across populations which are largely dependent on age.

While comparison across different longitudinal designs is one method of examining cohort effects, the Seattle Longitudinal Study achieved similar comparisons in a single study through the use of sampling with replacement (Schaie, 1994). In addition to repeated assessment of the initial sample, finding from the SLS have been meaningfully informed by the addition of new participants at each seven-year assessment. In all cases, participants have been drawn from the membership of a HMO group in the Seattle, Washington area and include a wide variety of professionals (from white- and blue-collar jobs) and their family members. Despite this limited commonality, each assessment group has included participants reflecting a wide range of ages.

Chief among the findings of the SLS is the presence of substantial generational differences across the six latent constructs according to participants' birth year. In other words, it's not only the case that participants' intellectual abilities vary by age but they also vary differentially from one generational cohort to the next. While several factors have been proposed to explain this effect (Flynn, 1987, 1999, 2000), correlational data from the SLS suggest that improvements and exposure to formal education are explanatory factors. In any case, the SLS highlights the unique power of longitudinal studies by suggesting that prior cross-sectional studies which explored age-related declines in cognitive ability may inaccurately estimate the degree of decline due to cohort differences (Schaie, 1994).

Among more recent longitudinal research, the Study of Mathematically Precocious Youth (SMPY) was begun by Stanley in 1971 and continued by Benbow, Lubinski, and their collaborators (Benbow et al., 1996) with the intent of identifying and addressing the educational needs of mathematically

gifted children. Though the scope of the study was later broadened slightly to include the needs of children who are gifted in other domains (Lubinski & Benbow, 2006), SMPY remains distinguished by the depth with which it has explored the relationship between the ability, temperament and interests of uniquely gifted children. Assessment is ongoing, but findings from SMPY will undoubtedly inform recent efforts to encourage greater interest among students in science, technology, engineering and mathematics (the "STEM" areas).

Brief with-in subject studies. The process of tracking subjects over long periods is both the primary advantage of longitudinal studies and the primary reason why they are not more widely implemented. Not only is it more costly and arduous to maintain contact with participants after the initial phase of data collection, but longitudinal designs seldom produce meaningful findings over a short time horizon (Costa & McCrae, 1992). One means of mitigating this aspect of longitudinal design is to limit the duration of the study and/or increase the frequency of data collection.

When the duration of study and frequency of data collection are drastically changed, as occurs in brief with-in subject studies, the resulting design may no longer appear longitudinal in nature (though still clearly distinct from cross-sectional). Studies of this type assess participants at very short intervals for a period of days or weeks and are used to explore the ways that behavior is effected by transient affective states, motivational pressures and diurnal rhythms. Of course, these designs cannot assess the long-term stability of attributes like typical longitudinal studies but this trade-off is acceptable when studying fine-grained behavioral patterns that are often lost between the infrequent measurement intervals of long-range studies.

Historically, experiments of this nature were restricted to the use of diary formats and suffered from issues related to data quality as a result. Fortunately, the introduction of several new technologies in recent years has helped to increase the ease of using this methodology. While cell phones are the most ubiquitous form of these technology, the list includes a broad array of self-tracking tools capable of measuring an increasing number of behavioral and interpersonal activities.

Of course the use of these technologies with longitudinal designs of longer durations is possible as well, but there are limits to the participants' willingness to devote their free time to academic research. While some existing technologies are able to collect and upload data via the Internet with minimal human involvement, the most germane data typically requires a degree of self-reflection on behalf of the participant. In this respect, long-term studies with high frequencies of data collection are not likely to employ current personality measures.

Nevertheless, the implementation of new data collection *technologies* will almost certainly influence the evolution of data collection *techniques*, and there is reason to believe this will be especially true in relation to brief within-subject designs. One hopes that further innovative development of these technologies will lead to exciting advances in personality research.

Methods of analysis

If Data = Model + Residual, the fundamental question of analysis is how to estimate the model? This depends, of course, on what the model is, but in general the method is to use the appropriate computational tool, whether this is a graphical description of the data or multi-wave, multi-level latent class analysis. For almost all problems facing the individual difference researcher, the appropriate computations can be done in the open source statistical system, R (R Development Core Team, 2009). Developed by a dedicated group of excellent statisticians, R has become the *lingua franca* of statistics and is becoming more used within psychology. In addition to the basic core R program which is freely available for download from the web, there are more than 2,000 specialized packages developed for different applications. A growing number of these packages are devoted to the problems of data analysis faced by the individual differences researcher (e.g., the *psych* package by Revelle, 2010). R is not only free but is also very powerful, it is the statistics system of choice for individual differences research.

Summary statistics and the problem of scaling

The most simple model of data is just the central tendency. But depending upon distributional properties such as skew, the two most common estimates (mean and median) can give drastically dif-

ferent values. Consider the case of family income in the United States according to the U.S. Census from 2008. Although mean family income was $66,570, median income was just $48,060. Any analysis using income as a covariate needs to take into account its log-normal characteristics. Besides offering graphical tools to detect such skewness, R has many ways to transform the data to produce "better behaved" data.

Non-linearities of the relationship between the latent variable of interest and the observed variable can lead to "fan-fold" interactions between ability and experimental manipulations (or just time) that suggest that individuals with higher initial scores change more or less than individuals with initially lower scores. Consider the hypothetical effect of one year of college upon writing and mathematics performance. Writing scores at one university go from 31 to 70 for an increase of 39 points but at another the scores go from 1 to 7 for an increase of 6 points. Most people would interpret this interaction (a gain of 39 versus a gain of 6 points) to reflect either the quality of instruction or the quality and motivation of the students. But when the same schools show that math scores at the first university improve just 6 points from 93 to 99 while going up 39 points (from 30 to 69) at the the other school, they interpret this change as representing a ceiling effect for the math test. But this interaction is exactly the same (although reversed) as the previous one. Such interactions due to the properties of the scale are also called *floor* and *ceiling* effects than can be eliminated with the appropriate monotone transformation. Unfortunately, these tend to be applied only if the interaction goes against expectation (Revelle, 2007).

The correlation coefficient and its nearest relatives

Sir Francis Galton may be credited with developing the theory of the correlation coefficient in his paper on "co-relations and their measurement" (Galton, 1888) which followed his paper (Galton, 1886) discussing the "the coefficient of reversion". Although the correlation was originally found by graphically fitting slopes to the medians for different values of the predictor (Galton, 1888), Pearson (1896) introduced the correlation coefficient bearing his name as the average cross product (the *covariance*) of standard scores

$$r_{xy} = Cov_{z_x z_y} = Cov_{\frac{x}{\sigma_x}\frac{y}{\sigma_y}} = \frac{Cov_{xy}}{\sigma_x \sigma_y} \tag{3}$$

and then Spearman (1904b) introduced the formula to psychologists in terms of deviation scores

$$r = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}. \tag{4}$$

It is Equation 4 that is most useful for seeing the relationship between the *Pearson Product Moment Correlation Coefficient* and a number of other measures of correlation (Table 1). When the data are continuous, r is known as a Pearson r. If the data are expressed in ranks, then this is just the Spearman *rho*. If X is dichotomous and Y continuous, the resulting correlation is known as a point bi-serial. If both X and Y are dichotomous, the correlation is known as Phi (ϕ). All of these use the same formula, although there are shortcuts that used to be used. Three additional correlation coefficients are listed which with the assumption of bivariate normality are equivalent to a Pearson r.

Researchers with an experimental bent tend to report seemingly different statistical estimates of the effect of one variable upon another. These are, however, merely transformations of the Pearson r (Table 2). Useful reviews of the use of these and other ways of estimating *effect sizes* for meta-analysis include Rosnow et al. (2000) and the special issue of Psychological Methods devoted to effect sizes (Becker, 2003).

With an appreciation of the different forms of the correlation it is possible to analyze traditional data sets more appropriately and to reach important conclusions. In medicine and clinical psychology for example, diagnoses tend to be categorical (someone is depressed or not, someone has an anxiety disorder or not). Co-occurrence of both of these symptoms is called *comorbidity*. Diagnostic categories vary in their degree of comorbidity with other diagnostic categories. From the point of view of correlation, comorbidity is just a name applied to one cell in a four fold table. It is possible to analyze comorbidity rates by considering the probability of the separate diagnoses and the probability of the joint diagnosis. This gives the two by two table needed for a ϕ or r_{tet} correlation. For instance, given the base rates (pro-

Table 1
A number of correlations are Pearson r in different forms, or with particular assumptions. If $r = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$, *then depending upon the type of data being analyzed, a variety of correlations are found.*

Coefficient	symbol	X	Y	Assumptions
Pearson	r	continuous	continuous	
Spearman	rho (ρ)	ranks	ranks	
Point bi-serial	r_{pb}	dichotomous	continuous	
Phi	ϕ	dichotomous	dichotomous	
Bi-serial	r_{bis}	dichotomous	continuous	normality
Tetrachoric	r_{tet}	dichotomous	dichotomous	bivariate normality
Polychoric	r_{pc}	categorical	categorical	bivariate normality

Table 2
Alternative Estimates of effect size. Using the correlation as a scale free estimate of effect size allows for combining experimental and correlational data in a metric that is directly interpretable as the effect of a standardized unit change in x leads to r change in standardized y.

Statistic	Estimate	r equivalent	as a function of r
Pearson correlation	$r_{xy} = \frac{C_{xy}}{\sigma_x \sigma_y}$	r_{xy}	
Regression	$b_{y.x} = \frac{C_{xy}}{\sigma_x^2}$	$r = b_{y.x} \frac{\sigma_y}{\sigma_x}$	$b_{y.x} = r \frac{\sigma_x}{\sigma_y}$
Cohen's d	$d = \frac{X_1 - X_2}{\sigma_x}$	$r = \frac{d}{\sqrt{d^2+4}}$	$d = \frac{2r}{\sqrt{1-r^2}}$
Hedge's g	$g = \frac{X_1 - X_2}{s_x}$	$r = \frac{g}{\sqrt{g^2+4(df/N)}}$	$g = \frac{2r\sqrt{df/N}}{\sqrt{1-r^2}}$
t - test	$t = 2d\sqrt{df}$	$r = \sqrt{t^2/(t^2+df)}$	$t = \sqrt{\frac{r^2 df}{1-r^2}}$
F-test	$F = 4d^2 df$	$r = \sqrt{F/(F+df)}$	$F = \frac{r^2 df}{1-r^2}$
Chi Square		$r = \sqrt{\chi^2/n}$	$\chi^2 = r^2 n$
Odds ratio	$d = \frac{ln(OR)}{1.81}$	$r = \frac{ln(OR)}{1.81\sqrt{(ln(OR)/1.81)^2+4}}$	$ln(OR) = \frac{3.62r}{\sqrt{1-r^2}}$
$r_{equivalent}$	r with probability p	$r = r_{equivalent}$	

portions) of two diagnostic categories (e.g., anxiety= .2 and depression =.15) and their co-occurence (comorbidity, e.g., .1), it is straightforward to find the tetrachoric correlation between the two diagnoses (.75). By using this basic fact, Krueger (2002) converted the comorbidities of various mental disorders to a matrix of tetrachoric correlations suitable for factor analysis and was able to argue for a two dimensional structure (internalizing and externalizing disorders) for a broad set of personality disorders.

Multiple R and the General Linear Model

A straight forward generalization of bivariate correlation and regression is the problem of multiple predictor variables and multiple correlation (Pearson, 1901). The problem is one of distinguishing between the *direct effect* of a predictor from the *total effect*. The total effect is the observed correlation, but the direct effect removes the effect of the other, correlated predictors. For a data matrix ${}_N X_n$ of N observations and n predictor variables and one criterion variable, y, if each of the predic-

tor variables ($x_1 \ldots x_n$) relates to y with correlations $r_{xy} = r_{x_1 y} \ldots r_{x_n y}$ and the x variables are themselves intercorrelated with correlation matrix R, then the predicted values of y ($\hat{y}$) are

$$\hat{y} = \beta X = r_{xy} R^{-1} X. \tag{5}$$

If the predictor set $x_i, \ldots, x_n$ are uncorrelated, then each separate variable makes a unique contribution to the dependent variable, y, and R^2, the amount of variance accounted for in y, is the sum of the individual r_{iy}^2. Unfortunately, most predictors are correlated, and the βs found in Equation 5 are less than the original correlations and since

$$R^2 = \sum \beta_i r_{x_i y} = \beta' r_{xy}$$

the R^2 will be less as the predictors become more correlated. An interesting, but unusual case, is that of *suppression* where a predictor, x_s does not relate to the criterion, y, but does relate to the other predictors. In this case x_s still is useful because it removes the variance in the other predictors not associated with the criterion. This leads to an interesting research problem for not only do we need to look for predictor of our criterion variable, we also need to look for non-predictors that predict the predictors!

The predictor set can be made up of any combination of variables, including the products or powers of the original variables. The products (especially when mean centered) represent the *interactions* of predictors (Cohen et al., 2003; Judd et al., 2009). Basic regression, multiple regression and graphic displays of residuals are all available in R using the `lm` or `glm` functions. The latter considers the case when the dependent (criterion) variable is dichotomous, such as success or failure (*logistic regression*), or discrete count data such as number of days missing school or number of times married (*Poisson, quasi-Poisson, and negative binomial regression*).

Spurious correlations

Although viewing the correlation coefficient as perhaps his greatest accomplishment, Pearson (1910) listed a number of sources of *spurious correlations* (Aldrich, 1995). These are challenges to all kinds of correlation, simple as well as multiple. Among these is the problem of ratios and of sums, and of correlations induced by mixing different groups. For the first problem, if two variables are expressed as ratios of a third variable, they will necessarily be correlated with each other. A related problem is when scores are forced to add up to a constant (i.e., they are *ipsatized*). In this case, even k uncorrelated variables will have a correlation of -1/(k-1) if they are ipsatized. As shown by Romer & Revelle (1984), the forced ipsatization of behavior ratings done by Shweder & D'Andrade (1980) led to the false claim of systematic distortion in interpersonal perception.

If data are pooled across groups, the overall correlation can be very different than the pooled within group correlation. Recognized as a problem since Yule (1912), *Simpsons paradox* (Simpson, 1951) was seen when sex discrimination in admissions was reported at the University of California, Berkeley. In 1973, UCB admitted about 44% of male applicants but, only about 35% of the females. What seems to be obvious sex discrimination in admissions became a paper in *Science* when it was discovered that the individual departments, if discriminating at all, discriminated in favor of women (Bickel et al., 1975). The women were applying to the departments which admitted fewer applicants as a percentage of applicants

The human eye and brain are superb pattern detectors. Using graphical displays rather than numeric tables helps detect strange relationships in one's data that are due to various artifact (Anscombe, 1973; Wainer, 1976; Wainer & Thissen, 1981). In a comparison of many statistical procedures to detect the underlying correlation in the presence of noise, the most robust estimator (least sensitive to noise and most sensitive to the underlying correlation) was the pooled estimates of a set of students trained to look at scatter plots (Wainer & Thissen, 1979).

Data quality: Reliability

The correlation of two variables is an index of the degree that variability in one is associated with variability in the other. It is not an index of causality, nor does it consider the quality of measurement of either variable. For X may directly cause Y, Y may directly cause X, or both may be caused by an unobserved third variable, Z. In addition, observed scores X and Y are probably not perfect representations of the constructs both are thought to measure.

Thinking back to Equation 1, the measure of X reflects a model of X as well as error in measurement. This realization led Spearman (1904b) to develop the basic concepts of reliability theory. He was the first psychologist to recognize that observed correlations are attenuated from the true correlation if the observations contain error.

> Now, suppose that we wish to ascertain the correspondence between a series of values, p, and another series, q. By practical observation we evidently do not obtain the true objective values, p and q, but only approximations which we will call p' and q'. Obviously, p' is less closely connected with q', than is p with q, for the first pair only correspond at all by the intermediation of the second pair; the real correspondence between p and q, shortly r_{pq} has been "attenuated" into $r_{p'q'}$ (Spearman, 1904b, p 90).

To Spearman, the reliability of a test, p', was the correlation with one just like it, p" (a parallel test). The problem of how to find test reliability has bedeviled psychometricians for more than 100 years (Spearman, 1904b), (Spearman, 1910), (Brown, 1910), (Guttman, 1945), Cronbach (1951), and we can only hope that we are coming to a solution (McDonald, 1999; Revelle & Zinbarg, 2009; Sijtsma, 2009).

Classical Test Theory. The solutions to the reliability question in classical test theory (Lord & Novick, 1968; McDonald, 1999) were extensions of the original suggestion by Spearman (1904b) for parallel tests. If estimated with two or more tests, the reliability of the composite is a function of the number of tests going into the composite (Brown, 1910; Spearman, 1910). Guttman (1945), although arguing that reliability was only meaningful over time, proposed six different ways of estimating reliability. One of these six (λ_3) was discussed later by Cronbach (1951) as *coefficient* α. Although routinely dismissed as an inappropriate estimate of reliability (Cronbach & Shavelson, 2004; McDonald, 1999; Revelle, 1979; Sijtsma, 2009; Zinbarg et al., 2005), α remains the most reported estimate of reliability. But α is always less than or equal to the true reliability (Guttman, 1945; Sijtsma, 2009) and is a poor way of assessing the homogeneity of a test. A test can have a substantial α even though the test measures two unrelated concepts (McDonald, 1999; Revelle, 1979; Revelle & Zinbarg, 2009). With the use of the `omega` function in the *psych* package, the two estimates developed by McDonald (1999), ω_h and ω_t are now easily calculated. ω_h (omega hierarchical) is the amount of variance that a general factor accounts for in a test and ω_t is the total amount of reliable variance in a test (McDonald, 1999; Revelle & Zinbarg, 2009). $\omega_h \leq \alpha \leq \omega_t$ and only in the case of a purely one factor test with equal item correlations will they be equal.

In addition to measures of reliability assessed using measures of a test's homogeneity, reliability is also of concern when measuring the same trait twice over an extended period of time. But such test-retest reliability or stability is not necessarily good for all measures. When assessing ability or a personality trait such as extraversion, test-retest reliability over extended periods of time is a sign of a stable trait. That IQ scores at age 11 correlate .66 with IQ scores at age 80 is remarkable and shows the stability of IQ (Deary et al., 2004). It is important to recognize that reliability is a rank order concept and that even with a perfect test-retest correlation, all the scores could have increased or decreased drastically. High test-retest reliability is not necessarily a good thing: to find a high test-retest of a measure of mood over a few days would imply that it is not a mood test, but rather a test of trait affectivity. That raters give similar ratings as other panel members on a selection board (Goldberg, 1966) is a sign of inter-rater reliability, a global measure of which can be found by using the Intra-Class Correlation (Shrout & Fleiss, 1979).

The intraclass correlation expresses the reliability of ratings in terms of components of variance associated with raters, targets, and their interactions and can be extended to other domains. That is, the analysis of variance approach to the measurement of reliability focuses on the relevant facets in an experimental design. If ratings are nested within teachers whom are nested within schools, and are given at different times, then all of these terms and their interactions are sources of variance in the ratings. First do an analysis of variance in a *generalizability study* to identify the variance components. Then determine which variance components are relevant for the application in the *decision study* in which

one is trying to use the measure (Cronbach et al., 1972). Similarly, the components of variance associated with parts of a test can be analyzed in terms of the generalizability of the entire test.

Item Response Theory: the new psychometrics. Classic psychometrics treats items as random replicates and models the total score. As such, reliability of measurement is a between person concept that does not allow a unique specification of the amount of error for each individual. Reliability is enhanced if the test variance goes up, and is meaningless for a single individual. The "new psychometrics" (Embretson & Hershberger, 1999), on the other hand, considers the information in each item and thus is able to talk about the precision of estimate for a score for a single person. Primary advantages of IRT procedures are that they can identify items that have *differential item functioning* (DIF) in different groups, test items can be formed into tests *tailored* for specific ability groups, and tests can be made *adaptive*. This ability to tailor a test to a particular difficulty level, and even more importantly, adaptively give items to reflect prior response patterns is one of the great strengths of IRT. For with a suitable item bank of many items, this allows researchers to give fewer items to any particular subject to obtain the same level of precision possible when using classical test methods. Examples of using IRT in clinical assessments include everything from measuring ease of breathing in cardiac patients to assessing psychopathology in the clinic (Reise & Waller, 2009). There has been an explosion of handbooks (Linden & Hambleton, 1997) and textbooks (Bond & Fox, 2007; Embretson, 1996; Embretson & Reise, 2000) on IRT and now, with R it is easy to do. However, to counter some of the enthusiasm for IRT, McDonald (1999) and Zickar & Broadfoot (2009) suggest that classical test theory is still alive and well and worth using for many applications. In most cases, the correlations of IRT and classical estimates are very high and perhaps the primary advantage of IRT modeling is the realization that observed responses are not linearly related to the latent trait being assessed.

Data usefulness: Validity

That a test or a judge gives the same value for a person over time is nice, but what is more important is do they give the right answer? Unfortunately, this is a much harder question to answer than is the test reliable. For what is the right answer? (Shooting an arrow into the same part of a target is reliability, hitting the bull's eye is validity, but this requires having a target.) Assessing validity requires having a criterion. This was the chief problem when selecting spies for the Office of Strategic Services (OSS Assessment Staff, 1948) as well as the selection of Peace Corps Volunteers (Wiggins, 1973), both classics in assessment, and both suffering from an unclear criterion. If the criterion is fuzzy, validity will necessarily be low.

With the focus on data as model plus residual, validity can be said to be measured by how well the model fits, compared to other models, and compared to what we would expect by chance. We prefer to have models using fewer parameters and not to be "multiplying entities beyond necessity"[2]. This implies there is not one validity, but rather a process of validation. Is a model useful? Is a model more useful than others? Is there a more simple model that does almost as well? This has become the domain of latent variable modeling.

Latent variable modeling

Spearman (1904b) recognized that the observed variable is befuddled with error (Equation 2) and that the underlying latent (or unobserved) score should be modeled when correcting correlations for unreliability. By *disattenuating* correlations, he hoped to study the underlying mechanisms. This switch from observed to latent variables was the basis for factor analysis and the search for a general factor of intelligence (Spearman, 1904a).

Factor analysis, Components Analysis, Cluster Analysis, Multidimensional scaling

Classical test theory is a model of how multiple items all measure a single latent trait. By knowing the latent variable and the resulting correlations of items with that latent variable, it is possible to perfectly predict the covariances between the items by taking the product of the respective correlations with the latent variable. This is the model known

[2] Although this dictum is probably neither original with William of Ockham nor directly stated by him (Thorburn, 1918), Ockham's razor remains a fundamental principal of science.

as a single factor. If all the items in a correlation matrix, R, are measures of latent variable, F, then the correlations can be modeled as

$$R = FF' + U^2 \tag{6}$$

where F is a vector (a one dimension matrix) of correlations of the variables with the latent factor, and U^2 is a diagonal matrix of residuals.

Even when generalizing this to more than one factor, Equation 6 remains the same matrix equation. Equation 6 when expressed in terms of single correlations, the elements of R, becomes for $i \neq j$

$$r_{ij} = \sum_{k=1}^{c} f_{ik} f_{jk} \tag{7}$$

that is, the correlation between any two variables is the sum of the products of their respective factor loadings on c factors.

Equation 6 is expressed in matrix algebra and is (with modern computational techniques) a very simple problem. As originally developed in terms of operations on tables of correlations (e.g., Equation 7) this was a difficult problem with one factor and an extremely difficult problem with more than one factor. However, with the introduction of matrix algebra to psychologists in the 1930s, Thurstone (1935) and others were able to exploit the power of matrix algebra (Bock, 2007). Recognizing that factor analysis (FA) was just a statistical model fitting problem and that goodness of fit statistics could be applied to the resulting solutions (Lawley & Maxwell, 1963) made factor analysis somewhat more respectable. The advent of powerful and readily available computers and computer algorithms to do factor analysis has led to much more frequent use of this powerful modeling technique.

Factor analysis models the observed patterns of correlations between the variables as the sum of the products of factors. At the structural level, this is just a problem of solving a set of simultaneous equations and (roughly speaking) if there are more correlations than unobserved factor loadings, the model is defined. Models with more or less factors can be compared in terms of how well they capture the original covariance or correlation matrix. However, because the factors are themselves unobservable, they can only be estimated. Thus, although completely defined at the structural level, factors are undefined at the level of the data.

This indeterminacy has led some to argue against factor analysis and in favor of principal components analysis (PCA). PCA forms linear sums of the observed variables to maximize the variance accounted for by successive components. These components, since they are linear sums of the observed variables, are completely determined. But the components, by summing the observed data, are no more parsimonious than the original data. If, however, just the first c components are extracted, then they are the best set of c independent linear sums to describe the data. Both factors and components have the same goal, to describe the original data and the original correlation matrix. Factor analysis models the off-diagonal elements (the common part) of the correlation matrix, while components model the entire correlation matrix. Although the two models are conceptually very different, and will produce very different results when examining the structure of a few ($< 20 - 30$) variables, they are unfortunately frequently confused, particularly by some of the major commercial statistical packages. The models are different and should not be seen as interchangeable.

Exploratory Factor Analysis (EFA) is used to find the structure of correlation matrices where items/tests are allowed to freely correlate with all factors. Rotations towards *simple structure* attempt to reduce the complexity of the solution and to make for more easily interpretable results. The factors as extracted from a EFA and the components as extracted from a PCA are independent. But it they are transformed to give them a simple structure where each item has a high correlation on one or only a few factors or components, then the factors/components probably will become correlated (oblique). What is the best transformation and how best to determine the optimal number of factors remains a point of debate although there is almost uniform agreement among psychometricians that number of factors with eigen values greater than one is the worst rule for determining the number of factors. This is, unfortunately, the default for many commercial programs.

A model which uses some of the logic of factor analysis but differs from EFA is cluster analysis. Hierarchical clustering algorithms (e.g., ICLUST, Revelle, 1979) combine similar pairs of items into clusters and hierarchically combine clusters until

some criteria (e.g, β or the worst split half reliability) fails to increase. ICLUST, as implement in R has proved useful in forming reliable and independent scales in an easily understood manner (Cooksey & Soutar, 2006; Markon, 2010).

An alternative data reduction and description technique that can produce drastically different solutions from FA or PCA is multidimensional scaling (MDS). MDS is also a fitting procedure, but when working with a correlation matrix, rather than treat the correlations as deviating from zero, MDS tries to minimize the deviations of the correlations from each other. That is to say, it fits the correlation matrix after removing the average correlation. The resulting solutions, particularly when the data have a general factor (e.g., ability tests) represent how different tests are from the average test, rather than how different correlations are from zero. This can be particularly useful when examining the microstructure of a battery of highly correlated tests.

Structural Equation modeling

Structural Equation Modeling (SEM) combines basic regression techniques with factor analysis modeling of the measurement of variables (Loehlin, 2004). Essentially, it is regression analysis applied to the dis-attenuated covariance matrix. In the modeling tradition it forces one to specify a model and then provides statistical estimates of fit that can be compared to alternative models. The power of SEM is that complex developmental growth models (McArdle, 2009), or hierarchical models of ability (Horn & McArdle, 2007) can be tested against alternative models. Examples applied to personality measurement include a Multi-trait Multi method analysis of the Big 5 (Biesanz & West, 2004). Perhaps a disadvantage of the ease of running SEM programs, is that some users are misled about the strength of their results. Because of the tendency to draw SEM path models with directional arrows, some users of SEM techniques mistakenly believe that they are testing causal models but are disabused of this when they realize that the models fit equally well when the "causal" direction is reversed. Other users fail to realize that a good model fit does not confirm a model and that it is necessary to consider fits of the multiplicity of alternative models.

Multi level modeling

The correlation within groups or individuals is not the same as the correlation between groups or individuals. What appears to be a strong relationship across groups can vanish when considering the individual within groups (Robinson, 1950; Yule, 1912). What had been seen as a challenge is now treated using the techniques of multilevel modeling. The use of multi-level modeling techniques (also known as Hierarchical Linear Models or multi-level Random Coefficient models) disentangle the effects of individuals from other, grouping effects in everything from developmental growth curve studies to studies of organizational effectiveness (Bliese et al., 2007). The clear two-three dimensional structure of affect as assessed between individuals (Rafaeli & Revelle, 2006) differs from individual to individual in terms of the patterning of affect experience overtime within individuals Rafaeli et al. (2007). What appear to be systematic effects of birth order on intelligence disappear when modeled within families (Wichman et al., 2006).

Computer modeling

Although hard to tell from reading most of literature in differential psychology, not all theories are tested by data analyzed using the general linear model. Some theories make predictions that are best tested using computer simulations. The theories are tested for reasonableness of results rather than fits to observations of the behavior of living subjects. The *Dynamics of Action* (Atkinson & Birch, 1970) and its reparameterization as the *Cues-Tendency-Action* (CTA) model (Fua et al., 2010; Revelle, 1986) predict dynamic patterning of behavior that is a non-linear consequence of the initial parameters. Connectionist models of personality (Read et al., 2010) or computational models of individual differences in reinforcement sensitivity (Pickering, 2008) make similar non-linear predictions that show the power of a few basic parameters in producing wide ranging variability in predicted outcome. Modeling is a method of research that has proven very powerful in fields ranging from climate research to evolutionary biology to cognitive psychology. With the ease of use of modeling software we can expect modeling to become a more common research method in differential psychology.

Conclusion

Differential Psychology is an extremely broad area of study. We have reviewed the major themes of data collection and methods of data analysis with the recognition that each section is worthy of a chapter in its own right. The basic theme is that Data = Model + Residual and the researcher needs to decide what constitutes data, what is an appropriate model, and what is reasonable to leave as a residual for someone else to model. In terms of data collection we are limited only by our imagination. Although great progress has been made since Galton and Spearman, the problems of data analysis remain the same.

References

Aldrich, J. (1995). Correlations genuine and spurious in Pearson and Yule. *Statistical Science*, *10*(4), 364-376.

Anderson, J., Lin, H., Treagust, D., Ross, S., & Yore, L. (2007). Using large-scale assessment datasets for research in science and mathematics education: Programme for International Student Assessment (PISA). *International Journal of Science and Mathematics Education*, *5*(4), 591–614.

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, *27*(1), 17–21.

Armstrong, T., & Olatunji, B. (2009). What they see is what you get: Eye tracking of attention in the anxiety disorders. *Psychological Science*, *23*(3).

Atkinson, J. W., & Birch, D. (1970). *The dynamics of action*. New York, N.Y.: John Wiley.

Baehr, E. K., Revelle, W., & Eastman, C. I. (2000). Individual differences in the phase and amplitude of the human circadian temperature rhythm: with an emphasis on morningness-eveningness. *Journal of Sleep Research*, *9*(2), 117–127.

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science*, *2*(4), 396-403.

Becker, B. J. (2003). Introduction to the special section on metric in meta-analysis. *Psychological Methods*, *8*(4), 403-405.

Benbow, C. P., Lubinski, D. J., & Stanley, J. C. (1996). *Intellectual talent: psychometric and social issues*. Baltimore: Johns Hopkins University Press.

Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, *187*(4175), 398-404.

Biesanz, J. C., & West, S. G. (2004). Towards understanding assessments of the Big Five: Multitrait-multimethod analyses of convergent and discriminant validity across measurement occasion and type of observer. *Journal of Personality*, *72*(4), 845–876.

Bliese, P. D., Chan, D., & Ployhart, R. E. (2007). Multilevel methods: Future directions in measurement, longitudinal analyses, and nonnormal outcomes. *Organizational Research Methods*, *10*(4), 551-563.

Block, J. (1971). *Lives through time*. Berkeley: Bancroft Books.

Block, J. (1993). Studying personality the long way. In D. C. Funder, R. Parke, C. Tomlinson-Keasey, & K. Widaman (Eds.), *Studying lives through time: personality and development* (pp. 9–41). Washington, D.C.: American Psychological Association.

Bock, R. D. (2007). Rethinking thurstone. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (p. 35-45). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Bond, T., & Fox, C. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ, US: Lawrence Erlbaum.

Borkenau, P., Mauer, N., Riemann, R., Spinath, F. M., & Angleitner, A. (2004). Thin slices of behavior as cues of personality and intelligence. *Journal of Personality and Social Psychology*, *86*(4), 599-614.

Borkenau, P., Riemann, R., Angleitner, A., & Spinath, F. M. (2001). Genetic and environmental influences on observed personality: Evidence from the german observational study of adult twins. *Journal of Personality and Social Psychology*, *80*(4), 655-668.

Born, W. K., Revelle, W., & Pinto, L. H. (2002). Improving biology performance with workshop groups. *Journal of Science Education and Technology*, *11*(4), 347-365.

Box, G. E. P. (1979). Some problems of statistics and everyday life. *Journal of the American Statistical Association*, *74*(365), 1–4.

Broadhurst, P. (1975, 10 29). The maudsley reactive and nonreactive strains of rats: A survey. *Behavior Genetics*, *5*(4), 299–319.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*(3), 296-322.

Buchanan, T., Johnson, J. A., & Goldberg, L. R. (2005). Implementing a five-factor personality inventory for use on the internet. *European Journal of Psychological Assessment*, *21*(2), 115-127.

Burisch, M. (1984). Approaches to personality inventory construction. *American Psychologist*, *39*(3), 214-227.

Burkhauser, R. V., & Lillard, D. R. (2007). The expanded Cross-National Equivalent File: HILDA joins its international peers. *Australian Economic Review*, *40*(2), 208–215.

Canli, T. (2004). Functional brain mapping of extraversion and neuroticism: learning from individual differences in emotion processing. *Journal of Personality*, *72*(6), 1105-1132.

Cattell, R. B. (1946). Personality structure and measurement. I. The operational determination of trait unities. *British Journal of Psychology*, *36*, 88-102.

Cattell, R. B. (1966). The data box: Its ordering of total resources in terms of possible relational systems. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology* (p. 67-128). Chicago: Rand-McNally.

Cattell, R. B., & Stice, G. (1957). *Handbook for the sixteen personality factor questionnaire.* Champaign, Ill.: Institute for Ability and Personality Testing.

Chernyshenko, O. S., Stark, S., Drasgow, F., & Roberts, B. W. (2007). Constructing personality scales under the assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment*, *19*(1), 88-106.

Chida, Y., & Steptoe, A. (2009). Cortisol awakening response and psychosocial factors: a systematic review and meta-analysis. *Biological Psychology*, *80*(3), 265-278.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309 - 319.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed ed.). Mahwah, N.J.: L. Erlbaum Associates.

Coleman, K., & Wilson, D. S. (1998). Shyness and boldness in pumpkinseed sunfish: individual differences are context-specific. *Animal Behaviour*, *56*(4), 927 - 936.

Cook, T., Campbell, D., & Day, A. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Houghton Mifflin Boston.

Cooksey, R., & Soutar, G. (2006). Coefficient beta and hierarchical item clustering - an analytical procedure for establishing and displaying the dimensionality and homogeneity of summated scales. *Organizational Research Methods*, *9*, 78-98.

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

Corr, P. J. (2008). *The reinforcement sensitivity theory of personality* (P. J. Corr, Ed.). Cambridge: Cambridge University Press.

Costa, P. T., & McCrae, R. R. (1985). *NEO PI professional manual*. Odessa, FL: Psychological Assessment Resources, Inc.

Costa, P. T., & McCrae, R. R. (1992). Multiple uses for longitudinal personality data. *European Journal of Personality*, *6*(2), 85–102.

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

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*, 671-684.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, *30*, 116-127.

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

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, *64*(3), 391-418.

Deary, I. J. (2009). Introduction to the special issue on cognitive epidemiology. *Intelligence*, *37*, 517–519.

Deary, I. J., Whiteman, M., Starr, J., Whalley, L., & Fox, H. (2004). The impact of childhood intelligence on later life: following up the Scottish mental surveys of 1932 and 1947. *Journal of Personality and Social Psychology*, *86*(1), 130–147.

DeYoung, C. G., Hirsh, J. B., Shane, M. S., Papademetris, X., Rajeevan, N., & Gray, J. R. (in press). Testing predictions from personality neuroscience. *Psychological Science*.

Diamond, L., & Otter-Henderson, K. D. (2007). Physiological measures. In R. Robins, C. R. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (p. 370-388). New York: The Guilford Press.

Digman, J. M. (1963). Principal dimensions of child personality as inferred from teachers' judgments. *Child Development*, *34*(1), 43–60. Available from http://www.jstor.org/stable/1126826

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology.*, *41*, 417-440.

Durbin, C. E., Hayden, E., Klein, D., & Olino, T. (2007). Stability of laboratory-assessed temperamental emotionality traits from ages 3 to 7. *Emotion*, *7*(2), 388–399.

Durbin, C. E., & Klein, D. N. (2006). 10-year stability of personality disorders among outpatients with mood disorders. *Journal of Abnormal Psychology.*, *115*(1), 75-84.

Durbin, C. E., Klein, D. N., Hayden, E. P., Buckley, M. E., & Moerk, K. C. (2005). Temperamental emotionality in preschoolers and parental mood disorders. *Journal of Abnormal Psychology*, *114*(1), 28-37.

Ekman, P., Friesen, W. V., & Hager, J. C. (1978). *Facial action coding system*. Palo Alto, CA: Consulting Psychologists Press.

Elder, G. H. (1998). The life course as developmental theory. *Child Development*, *69*(1), 1–12.

Elder, G. H., Shanahan, M., & Clipp, E. (1994). When war comes to men's lives: Life-course patterns in family, work, and health. *Psychology and Aging*, *9*(1), 5–16.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, *8*(4), 341-349.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*(3), 380 - 396.

Embretson, S. E., & Hershberger, S. L. (1999). *The new rules of measurement : what every psychologist and educator should know*. Mahwah, N.J.: L. Erlbaum Associates.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, N.J.: L. Erlbaum Associates.

Eysenck, H. J. (1966). Personality and experimental psychology. *Bulletin of the British Psychological Society*, *19*, 1-28.

Eysenck, H. J. (1967). *The biological basis of personality*. Springfield: Thomas.

Eysenck, H. J. (1997). Personality and experimental psychology: The unification of psychology and the possibility of a paradigm. *Journal of Personality and Social Psychology*, *73*(6), 1224-1237.

Eysenck, H. J., & Eysenck, S. B. G. (1968). *Manual for the Eysenck Personality Inventory*. San Diego, CA: Educational and Industrial Testing Service.

Fiske, D. W. (1949). Consistency of the factorial structures of personality ratings from different sources. *Journal of Abnormal and Social Psychology*, *44*, 329-344.

Fleeson, W., Malanos, A. B., & Achille, N. M. (2002). An intraindividual process approach to the relationship between extraversion and positive affect: Is acting extraverted as "good" as being extraverted? *Journal of Personality and Social Psychology*, *83*(6), 1409-1422.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*, *101*(2), 171-191.

Flynn, J. R. (1999). Searching for justice: The discovery of IQ gains over time. *American Psychologist*, *54*(1), 5–20.

Flynn, J. R. (2000). IQ gains and fluid g. *American Psychologist*, *55*(5), 543.

Fox, E., Russo, R., Bowles, R., & Dutton, K. (2001). Do threatening stimuli draw or hold visual attention in subclinical anxiety? *Journal of Experimental Psychology: General*, *130*(4), 681-700.

Friedman, H. S., Tucker, J. S., Schwartz, J. E., Tomlinson-Keasey, C., Martin, L. R., Wingard, D. L., et al. (1995). Psychosocial and behavioral predictors of longevity: The aging and death of the "termites". *American Psychologist*, *50*(2), 69 - 78.

Fua, K., Revelle, W., & Ortony, A. (2010). Modeling personality and individual differences: the approach-avoid-conflict triad. In *Cogsci 2010.*

Funder, D. C. (2001). Personality. *Annual Review of Psychology*, *52*, 197-221.

Funder, D. C., Furr, R., & Colvin, C. (2000). The Riverside Behavioral Q-sort: A tool for the description of social behavior. *Journal of Personality*, *68*(3), 451-489.

Furr, R., & Funder, D. (2007). Behavioral observation. *Handbook of research methods in personality psychology*, 273–291.

Gallagher, E. J., Viscoli, C. M., & Horwitz, R. I. (1993). The relationship of treatment adherence to the risk of death after myocardial infarction in women. *JAMA*, *270*(6), 742-744.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, *15*, 246-263.

Galton, F. (1888). Co-relations and their measurement. *Proceedings of the Royal Society. London Series*, *45*, 135-145.

Goldberg, L. R. (1966). Reliability of peace corps selection boards: A study of interjudge agreement before and after board discussions. *Journal of Applied Psychology*, *50*(5), 400 - 408.

Goldberg, L. R. (1972). Parameters of personality inventory construction and utilization: A comparison of prediction strategies and tactics. *Multivariate Behavioral Research Monographs. No 72-2*.

Goldberg, L. R. (1990). An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology*, *59*(6), 1216-1229.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, p. 7-28). Tilburg, The Netherlands: Tilburg University Press.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., et al. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, *40*(1), 84-96.

Gosling, S. D. (2001). From mice to men: What can we learn about personality from animal research? *Psychological Bulletin*, *127*(1), 45–86.

Gosling, S. D., & John, O. P. (1999). Personality dimensions in nonhuman animals: A cross-species review. *Current Directions in Psychological Science*, *8*(3), 69–75.

Gosling, S. D., & Vazire, S. (2002). Are we barking up the right tree? evaluating a comparative approach to personality. *Journal of Research in Personality*, *36*(6), 607-614.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, *59*(2), 93–104.

Gough, H. G. (1957). *Manual for the California psychological inventory.* Palo Alto, CA: Consulting Psychologists Press.

Gray, J. A. (1982). *Neuropsychological theory of anxiety: An investigation of the septal-hippocampal system*. Cambridge: Cambridge University Press.

Gray, J. A., & McNaughton, N. (2000). *The neuropsychology of anxiety: An enquiry into the functions of the septo-hippocampal system* (2nd ed.). Oxford: Oxford University Press.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(4), 255–282.

Habermas, T., & Bluck, S. (2000). Getting a life: The emergence of the life story in adolescence. *Psychological Bulletin*(126), 248-269.

Harmon-Jones, E., & Beer, J. S. (2009). *Methods in social neuroscience*. New York, NY US: Guilford Press.

Harvey, E. (1999). Short-term and long-term effects of early parental employment on children of the National Longitudinal Survey of Youth. *Developmental Psychology*, *35*(2), 445–459.

Hase, H. D., & Goldberg, L. R. (1967). Comparative validity of different strategies of constructing personality inventory scales. *Psychological Bulletin*, *67*(4), 231-248.

Hathaway, S., & McKinley, J. (1943). *Manual for administering and scoring the MMPI.* Minneapolis: University of Minnesota Press.

Hogan, J., Barrett, P., & Hogan, R. (2007). Personality measurement, faking, and employment selection. *Journal of Applied Psychology*, *92*(5), 1270 - 1285.

Hogan, R. (2007). *Personality and the fate of organizations*. ix, 167 pp. Mahwah, NJ: Lawrence Erlbaum Associates Publishers Lawrence Erlbaum Associates Publishers.

Hogan, R., & Kaiser, R. B. (2005). What we know about leadership. *Review of General Psychology*, *9*(2), 169 - 180.

Hogan, R., & Nicholson, R. A. (1988). The meaning of personality test scores. *American Psychologist*, *43*(8), 621 - 626.

Holland, J. L. (1959). A theory of vocational choice. *Journal of Counseling Psychology*, *6*(1), 35 - 45.

Holland, J. L. (1996). Exploring careers with a typology: What we have learned and some new directions. *American Psychologist*, *51*(4), 397 - 406.

Horn, J. L., & McArdle, J. J. (2007). Understanding human intelligence since Spearman. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: historical developments and future directions* (p. 205-247). Mahwah, N.J.: Erlbaum.

Horwitz, R. I., Viscoli, C. M., Donaldson, R. M., Murray, C. J., Ransohoff, D. F., Berkman, L., et al. (1990). Treatment adherence and risk of death after a myocardial infarction. *Lancet*, *336*(8714), 542-545.

Humphreys, M. S., & Revelle, W. (1984). Personality, motivation, and performance: A theory of the relationship between individual differences and information processing. *Psychological Review*, *91*(2), 153-184.

Hunt, E., & Wittmann, W. (2008). National intelligence and national prosperity. *Intelligence*, *36*(1), 1 - 9.

Irvine, J., Baker, B., Smith, J., Jandciu, S., Paquette, M., Cairns, J., et al. (1999). Poor adherence to placebo or amiodarone therapy predicts mortality: results from the CAMIAT study. *Psychosomatic Medicine*, *61*(4), 566-575.

Jackson, D. N. (1967). *Personality research form manual*. Research Psychologists Press.

Johnson, D. L., Wiebe, J. S., Gold, S. M., Andreasen, N. C., Hichwa, R. D., Watkins, G., et al. (1999). Cerebral blood flow and personality: A positron emission tomography study. *American Journal of Psychiatry*, *156*(2), 252-257.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, *39*(1), 103-129.

Judd, C. M., McClelland, G. H., & Ryan, C. S. (2009). *Data analysis: a model comparison approach* (2nd ed ed.). New York: Routledge.

Kelly, E. L., & Fiske, D. W. (1950). The prediction of success in the va training program in clinical psychology. *American Psychologist*, *5*(8), 395 - 406.

Krueger, R. F. (2002). Psychometric perspectives on comorbidity. In J. E. Helzer & J. J. Hudziak (Eds.), *Defining psychopathology in the 21st century: DSM-V and beyond* (p. 41-54). Washington, DC: American Psychiatric Publishing, Inc American Psychiatric Publishing, Inc.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the graduate record examinations: Implications for

graduate student selection and performance. *Psychological Bulletin*, *127*(1), 162 - 181.

Kuncel, N. R., Kuncel, N. R., Credé, M., & Thomas, L. L. (2007). A meta-analysis of the predictive validity of the Graduate Management Admission Test (GMAT) and undergraduate grade point average (UGPA) for graduate student academic performance. *The Academy of Management Learning and Education*, *6*(1), 51-68.

Lawley, D. N., & Maxwell, A. E. (1963). *Factor analysis as a statistical method*. London: Butterworths.

Linden, W. Van der, & Hambleton, R. (1997). *Handbook of modern item response theory*. Springer Verlag.

Loehlin, J. C. (2004). *Latent variable models: an introduction to factor, path, and structural equation analysis* (4th ed.). Mahwah, N.J.: L. Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley Pub. Co.

Lubinski, D., & Benbow, C. P. (2006). Study of mathematically precocious youth after 35 years: Uncovering antecedents for the development of math-science expertise. *Perspectives on Psychological Science*, *1*(4), 316–345.

Markon, K. E. (2010). Modeling psychopathology structure: a symptom-level analysis of axis i and ii disorders. *Psychological Medicine*, *40*, 273-288.

McAdams, D. P. (1993). *The stories we live by: Personal myths and the making of the self*. New York, NY: William Morrow & Co.

McAdams, D. P. (2008). Personal narratives and the life story. In *Handbook of personality: Theory and research* (3rd ed.). New York, NY: Guilford Press.

McAdams, D. P., Diamond, A., St. Aubin, E. de, & Mansfield, E. (1997). Stories of commitment: The psychosocial construction of generative lives. *Journal of Personality and Social Psychology*, *72*(3), 678-694.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology. Vol 60 Jan 2009, 577-605.*.

McArdle, J. J., & Bell, R. Q. (2000). Recent trends in modeling longitudinal data by latent growth curve methods. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multiple-group data: practical issues, applied approaches, and scientific examples* (p. 69-107). Mahwah, NJ: Erlbaum.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, N.J.: L. Erlbaum Associates.

McLean, K. C. (2005). Late adolescent identity development: Narrative meaning making and memory telling. *Developmental Psychology*, *41*(4), 683-691.

Mehl, M. R., & Pennebaker, J. W. (2003). The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of Personality and Social Psychology*, *84*(4), 857-870.

Mehl, M. R., Vazire, S., Ramirez-Esparza, N., Slatcher, R. B., & Pennebaker, J. W. (2007). Are women really more talkative than men? *Science*, *317*(5834), 82.

Mroczek, D. K. (2007). The analysis of longitudinal data in personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (p. 543-556). New York, NY: Guilford Press.

Mulaik, S. A. (2010). *Foundations of factor analysis* (2nd ed ed.). Boca Raton: CRC Press.

Nesselroade, J. R. (1984). Concepts of intraindividual variabiilty and change: impressions of Cattell's influence on lifespan developmental psycholgy. *Multivariate Behavioral Research*, *19*(2), 269-286.

Noftle, E. E., & Fleeson, W. (2010). Age differences in big five behavior averages and variabilities across the adult life span: Moving beyond retrospective, global summary accounts of personality. *Psychology and Aging*, *25*(1), 95-107.

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factors structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology*, *66*(6), 574-583.

Norman, W. T. (1969). "to see ourseles as ithers see us": Relations among self-perceptions, peer-perceptions, and expected peer-perceptions of personality attributes. *Multivariate Behavioral Research*, *4*(4), 417-443.

OSS Assessment Staff. (1948). *Assessment of men: Selection of personnel for the office of strategic services*. New York: Rinehart.

Pasupathi, M., & Hoyt, T. (2009). The development of narrative identity in late adolescence and emergent adulthood: The continued importance of listeners. *Developmental Psychology*, *45*(2), 558-574.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Causality: Models, reasoning, and inference xvi, 384 pp New York, NY, US: Cambridge University Press.

Pearson, K. P. (1896). Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia. *Philisopical Transactions of the Royal Society of London. Series A*, *187*, 254-318.

Pearson, K. P. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh and Dublin Philosophical Magazine and Journal*, *6*(2), 559-572.

Pearson, K. P. (1910). *The grammar of science* (3rd ed.). London: Adam and Charles Black.

Pennebaker, J. W., Mayne, T. J., & Francis, M. E. (1997). Linguistic predictors of adaptive bereavement. *Journal of Personality and Social Psychology*, *72*(4), 863-871.

Perkins, A. M., Kemp, S. E., & Corr, P. J. (2007). Fear and anxiety as separable emotions: An investigation of the revised reinforcement sensitivity theory of personality. *Emotion*, *7*(2), 252-261.

Pickering, A. D. (2008). Formal and computational models of reinforcement sensitivity theory. In P. J. Corr (Ed.), *The reinforcement sensivity theory.* Cambridge: Cambridge University Press.

R Development Core Team. (2009). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from http://www.R-project.org (ISBN 3-900051-07-0)

Rafaeli, E., & Revelle, W. (2006). A premature consensus: Are happiness and sadness truly opposite affects? *Motivation and Emotion*, *30*(1), 1-12.

Rafaeli, E., Rogers, G. M., & Revelle, W. (2007). Affective synchrony: Individual differences in mixed emotions. *Personality and Social Psychology Bulletin*, *33*(7), 915-932.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: reprinted in 1980 by The University of Chicago Press /Paedagogike Institut, Copenhagen.

Read, S. J., Monroe, B. M., Brownstein, A. L., Yang, Y., Chopra, G., & Miller, L. C. (2010). A neural network model of the structure and dynamics of human personality. *Psychological Review*, *117*(1), 61 - 92.

Reise, S., & Waller, N. (2009). Item response theory and clinical measurement. *Annual review of clinical psychology*, *5*, 27–48.

Revelle, W. (1979). Hierarchical cluster-analysis and the internal structure of tests. *Multivariate Behavioral Research*, *14*(1), 57-74.

Revelle, W. (1986). Motivation and efficiency of cognitive performance. In D. R. Brown & J. Veroff (Eds.), *Frontiers of motivational psychology: Essays in honor of J. W. Atkinson* (p. 105-131). New York: Springer.

Revelle, W. (1993). Individual differences in personality and motivation: 'non-cognitive' determinants of cognitive performance. In A. Baddeley & L. Weiskrantz (Eds.), *Attention: Selection, awareness, and control: A tribute to Donald Broadbent* (p. 346-373). New York, NY: Clarendon Press/Oxford University Press.

Revelle, W. (2007). Experimental approaches to the study of personality. In R. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology.* (p. 37-61). New York: Guilford.

Revelle, W. (2010). psych: Procedures for personality and psychological research (1.0-90 ed.) [Computer software manual]. Evanston. Available from http://personality-project.org/r (R package version 1.0-90)

Revelle, W., Amaral, P., & Turriff, S. (1976). Introversion-extraversion, time stress, and caffeine: the effect on verbal performance. *Science*, *192*, 149-150.

Revelle, W., Humphreys, M. S., Simon, L., & Gilliland, K. (1980). Interactive effect of personality, time of day, and caffeine - test of the arousal model. *Journal of Experimental Psychology General*, *109*(1), 1-31.

Revelle, W., & Oehleberg, K. (2008). Integrating experimental and observational personality research – the contributions of Hans Eysenck. *Journal of Personality*, *76*(6), 1387-1414.

Revelle, W., Wilt, J., & Rosenthal, A. (2010). Personality and cognition: The personality-cognition link. In A. Gruszka, G. Matthews, & B. Szymura (Eds.), *Handbook of individual differences in cognition: Attention, memory and executive control* (p. 27-49). Springer.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega and the glb: comments on Sijtsma. *Psychometrika*, *74*(1), 145-154.

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science*, *2*(4), 313-345.

Robins, R. W., Fraley, R. C., & Krueger, R. F. (2007). *Handbook of research methods in personality psychology*. Handbook of research methods in personality psychology. xiii, 719 pp. New York, NY: Guilford Press.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, *15*(3), 351–357.

Rocklin, T., & Revelle, W. (1981). The measurement of extraversion: A comparison of the Eysenck Personality Inventory and the Eysenck Personality Questionnaire. *British Journal of Social Psychology*, *20*(4), 279-284.

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist*, *65*(1), 1 - 12.

Romer, D., & Revelle, W. (1984). Personality traits: Fact or fiction? a critique of the Shweder and D'Andrade systematic distortion hypothesis. *Journal of Personality and Social Psychology*, *47*(5), 1028-1042.

Rosnow, R. L., Rosenthal, R., & Rubin, D. B. (2000). Contrasts and correlations in effect-size estimation. *Psychological Science*, *11*(6), 446-453.

Schaie, K. W. (1994). The course of adult intellectual development. *American Psychologist*, *49*(4), 304–313.

Schimmack, U., & Reisenzein, R. (2002). Experiencing activation: Energetic arousal and tense arousal are not mixtures of valence and activation. *Emotion*, *2*(4), 412-417.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420-428.

Shweder, R. A., & D'Andrade, R. G. (1980). The systematic distortion hypothesis. In R. A. Shweder (Ed.), *New directions for methodology of social and behavior science s* (p. 37-58). San Francisco: Jossey-Bass.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of cronbach's alpha. *Psychometrika*, *74*(1), 107-120.

Simms, L. J., & Watson, D. (2007). The construct validation approach to personality scale construction. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (p. 240-258). New York, NY: Guilford Press.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *13*(2), 238–241.

Singer, J. A. (2004). Narrative identity and meaning making across the adult lifespan: An introduction. *Journal of Personality*, *72*(3), 437-459.

Singer, J. A., & Blagov, P. (2004). The integrative function of narrative processing: Autobiographical memory, self-defining memories, and the life story of identity. In *The self and memory* (p. 117-138). New York, NY: Psychology Press.

Smillie, L., & Jackson, C. (2006). Functional impulsivity and reinforcement sensitivity theory. *Journal of Personality*, *74*(1), 47.

Spearman, C. (1904a). "general intelligence," objectively determined and measured. *American Journal of Psychology*, *15*(2), 201-292.

Spearman, C. (1904b). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72-101.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*(3), 271-295.

Strong, E. K. (1927). Vocational interest test. *Educational Record*, *8*(2), 107-121.

Terman, L. M., & Oden, M. (1947). *Genetic studies of genius*. Palo Alto, CA: Stanford University Press; Oxford University Press.

Thayer, R. E. (2000). *Mood*. (2000). Encyclopedia of psychology, Vol. 5. (pp. 294-295). Washington, DC ; New York, NY: American Psychological Association; Oxford University Press 508 pp.

Thorburn, W. M. (1918). The myth of occam's razor. *Mind*, *27*, 345-353.

Thurstone, L. L. (1935). *The vectors of mind: multiple-factor analysis for the isolation of primary traits*. Chicago: Univ. of Chicago Press.

Tupes, E. C., & Christal, R. E. (1961). *Recurrent personality factors based on trait ratings* (Tech. Rep. No. 61-97). Lackland Air Force Base: USAF ASD Technical Report.

Vale, J., & Vale, C. (1969). Individual differences and general laws in psychology: a reconciliation. *American Psychologist*, *24*(12), 1093–1108.

Vazire, S. (2006). Informant reports: A cheap, fast, and easy method for personality assessment. *Journal of Research in Personality*, *40*(5), 472-481.

Vazire, S., Gosling, S. D., Dickey, A. S., & Schapiro, S. J. (2007). Measuring personality in nonhuman animals. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (p. 190-206). New York, NY: Guilford Press.

Wacker, J., Chavanon, M.-L., & Stemmler, G. (2006). Investigating the dopaminergic basis of extraversion in humans: A multilevel approach. *Journal of Personality and Social Psychology*, *91*(1), 171-187.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, *83*(2), 213-217.

Wainer, H., & Thissen, D. (1979). On the Robustness of a Class of Naive Estimators. *Applied Psychological Measurement*, *3*(4), 543-551.

Wainer, H., & Thissen, D. (1981). Graphical data analysis. *Annual Review of Psychology*, *32*, 191 - 241.

Walker, R. N. (1967). Some temperament traits in children as viewed by their peers, their teachers, and themselves. *Monographs of the Society for Research in Child Development*, *32*(6), iii–36.

Watson, D. (2005). Rethinking the mood and anxiety disorders: a quantitative hierarchical model for DSM-V. *Journal of Abnormal Psychology*, *114*(4), 522.

Wichman, A., Rodgers, J., & MacCallum, R. (2006). A multilevel approach to the relationship between birth order and intelligence. *Personality and social psychology bulletin*, *32*(1), 117.

Wiggins, J. S. (1973). *Personality and prediction: principles of personality assessment*. Malabar, Fla.: R.E. Krieger Pub. Co.

Wilt, J., Oehleberg, K., & Revelle, W. (in press). Anxiety in personality. *Personality and Individual Differences*.

Yule, G. U. (1912). On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society*, *LXXV*, 579-652.

Zickar, M. J., & Broadfoot, A. A. (2009). The partial revival of a dead horse? comparing classical test theory and item response theory. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences.* (p. 37-59). New York, NY, US: Routledge/Taylor & Francis Group.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω_H): Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123-133.