The “Error” in Psychology: An Analysis of Quantitative and Qualitative Approaches in the Pursuit of Accuracy

by

Donna Tafreshi

MA, Simon Fraser University, 2014

BA, Simon Fraser University, 2011

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Psychology
Faculty of Arts and Social Sciences

SIMON FRASER UNIVERSITY

Fall 2018

Approval

Name:	**Donna Tafreshi**
Degree:	**Doctor of Philosophy**
Title:	**The "Error" in Psychology: An Analysis of Quantitative and Qualitative Approaches in the Pursuit of Accuracy**
Examining Committee:	**Chair:** Thomas Spalek Professor
	Kathleen Slaney Senior Supervisor Professor
	Timothy Racine Supervisor Professor
	Susan O'Neill Supervisor Professor
	Barbara Mitchell Internal Examiner Professor Departments of Sociology & Gerontology
	Christopher Green External Examiner Professor Department of Psychology York University
Date Defended/Approved:	October 12, 2018

Abstract

The concept of “error” is central to the development and use of statistical tools in psychology. Yet, little work has focused on elucidating its conceptual meanings and the potential implications for research practice. I explore the emergence of uses of the “error” concept within the field of psychology through a historical mapping of its uses from early observational astronomy, to the study of social statistics, and subsequently to its adoption under 20th century psychometrics. In so doing, I consider the philosophical foundations on which the concepts “error” and “true score” are built and the relevance of these foundations for its usages in psychology. Given the recent surge in interest in qualitative research methods in psychology, I also investigate whether a notion of “error” is relevant to qualitative research practice. In particular, I conduct a content analysis of usages of the term “reliability” within the qualitative methodological literature as a proxy for the concept of “error” within the qualitative research domain. Finally, I compare my explorations of discourse around quantitative and qualitative methods. I conclude that although researchers using methodological tools from these two traditions may hold opposing views on knowledge and truth, they also share a common aim of accuracy. Implications for research practice and education in psychology are discussed.

Keywords: measurement error; true scores; accuracy; certainty; reliability; quantitative and qualitative methods

Dedicated to the memory of my friend, Joshua Scott Patillo.

Acknowledgements

I thank Dr. Kate Slaney for her guidance and support as a supervisor, mentor, and friend. I also thank Dr. Tim Racine for his encouragement and wisdom throughout my graduate school career and I thank Dr. Susan O'Neill for her enthusiasm and knowledge on teaching qualitative methods.

I thank my partner, Minilik Joseph, for putting up with yet another graduate degree. Thank you for keeping me grounded and always being there for me, even during the toughest times. I thank my family for their unconditional support and I thank my parents for the sacrifices they made to allow me to pursue my education.

Table of Contents

List of Tables

List of Figures

Chapter 1.

Introduction

"Error" is a fundamental concept relevant to most, if not all, statistical tools used by psychologists today. It is an essential building block of reliability analysis, a basic assumption of which is that greater reliability indicates greater precision of a measurement or statistical estimate, and therefore, less error associated with that measurement or estimate. It is also a concept that is germane to the conceptualization of statistical models (e.g., fixed vs. random vs. mixed). Error has become especially relevant in areas of study where advanced statistical tools that allow for the modeling of nested systematic errors have become quite popular (e.g., multilevel modeling, hierarchical linear modeling, growth curve analysis). Moreover, how one conceptualizes error can have implications for the kinds of analyses used, as well as the kinds of inferences that can be justified based on those analyses. Thus, the concept of "error" plays a focal role in statistical methods and their application in psychological research. More importantly, perhaps, "error" plays a crucial role in science's pursuit for certainty in measurements or estimates. As Stigler (1986) noted, simply producing a measurement is not quite enough to satisfy the task of a scientist. "To serve the purposes of science the measurements must be susceptible to comparison," we must have a way of "expressing the uncertainty in their values and the inferential statements derived from them" (Stigler, 1986, p. 1). In other words, scientists want their measurements to be *accurate,* and *consistently* so. Over the course of the history of statistics and quantitative psychology, the concept of "error" has been central to theory and practice relevant to the pursuit of such accuracy. The first aim of the current project is to trace, chart, and analyze the conceptualization(s) of error throughout this history.

In methodology textbooks pertaining to psychology, "error" is commonly described as the difference between a central value (typically a "mean") and other values in a population of scores. At times, this population is described as representing infinite repeated scores of the *same* observation (i.e., intra-individual variability); whereas, at other times, it represents infinite repeated scores on a measurement obtained from *different* people (i.e., inter-individual variability). This distinction, between

intra- and inter- individual variability, is not always made explicit. Moreover, “error” appears to most commonly be defined computationally; much less common are discussions of its conceptual meaning. The term “error” itself is one that is found frequently in every day discourse. It is defined in both the Cambridge and Oxford online dictionaries as a “mistake.” However, several authors have clarified that *statistical* error does *not* represent a mistake in research practice (e.g., Thorndike, 1966; Stanley, 1971). Thus, it seems clear that statisticians, researchers, and psychologists more specifically, aim to use error in a technical, rather than every-day, sense. Indeed, in the most basic computational sense, error is simply a deviation “score.” Yet, what this deviation represents is less obvious. To my knowledge, no work within psychology or the philosophy of science has directly examined the conceptual meaning(s) of “error,” or the implications that it may have for psychological research. I will argue that clarification of the meaning(s) of the statistical concept of “error” is important to its use in psychological research practice. This involves, among other things, clarifying the distinction between errors at the intra- and inter- individual levels of measurement.

The second aim of the current project is to examine the statistical concept of “error” in relation to *qualitative* research methods in psychology (i.e., research based on non-numerical representations of data and non-statistical tools of analysis). Although not as popular as quantitative analysis, qualitative research has gained greater acceptance amongst North American psychologists in the past several years. For example, Division 5 of the American Psychological Association was recently renamed from “Measurement, Evaluation, and Statistics” to “Quantitative and Qualitative Methods” (see Gergen, Josselson, & Freeman, 2015). With a growing interest in qualitative research, the relations between quantitative and qualitative research have become increasingly pertinent. There have been debates within the social science methodological literature for some time now regarding the compatibility of quantitative and qualitative analysis, with many arguing that these two modes of research stem from fundamentally different philosophical presumptions (see Sale, Lohfeld, & Brazil, 2002). If this is the case, the concept of error may be tied strictly to “quantitative” ways of representing and analyzing psychological phenomena.

Nonetheless, some qualitative researchers have examined psychometric concepts related to “error” in light of qualitative methodology. For example, Norris (1997) and Wertz (1986) each attempted to outline a form of “reliability” that is applicable to

qualitative analysis. Although the statistical concept of "error" is not explicitly relevant to qualitative research, the ontological and epistemological implications of a given conceptualization of error may have bearing on qualitative research practice. For example, the notion of measurement error was originally introduced into the physical sciences based on a particular philosophical conception of phenomena under study. That is, it was believed that objects of study (e.g., planets) had a "true" existence that was outside of the influence of observers (Porter, 1986; Stigler, 1986). Thus, "error" was originally introduced as a concept denoting the difference between one's observation of an object and the object itself, as it "truly" exists. Such a conceptualization of error may be difficult to reconcile in qualitative research where it is common to take the perspective that phenomena do not exist independent of influence from observers. Of course, the conception and use of error within scientific practice has changed over the years. Yet, in modern times, there appears to be a lack of clarity regarding interpretations of the meaning of error and its relevance for the study of psychological phenomena. Given this, it is reasonable to ask: Are there senses of "error" that are relevant to both quantitative and qualitative research practice? Thus, the final aim of my current project is to compare the conceptualizations of error within quantitative and qualitative research methods (with a focus on the field of psychology).

Given the above considerations, the three overarching aims of the current project can be summarized as follows:

1. **To clarify the various ways in which "error" has been conceptualized in quantitative psychology and to understand the implications of these various conceptualizations for psychological research.** To this end, I conducted an analysis of the historical uses of the error concept within psychology, as well as in early statistics. My justification for taking a historical approach is twofold. First, the history of "error" is intricately related to the history of statistics. Error is a concept that was critical to the inception of the earliest statistical tools (e.g., the error law). Second, given that error is a concept central to statistics, it has been theorized about in explicit ways within the statistics and psychometrics literatures. Thus, taking a historical approach allowed me to identify important events in the history of the concept that have contributed to its conceptualization (and, perhaps also, have led to a multiplicity of conceptualizations).

2. **To examine whether the concept of "error" plays a role in qualitative methods (and to further explore what that role might be).** This was done through a thematic analysis of the qualitative

methodological literature with a focus on examining uses of "reliability."

3. **To compare the conceptualization(s) of error identified in aims 1 and 2 (within quantitative and qualitative research methods) and discuss implications.**

Chapter 2.

Error in Historical Perspective Part I: Early Uses of Error in Statistics and the Study of Social and Psychological Phenomena

In this first section, I trace important events prior to the 20th century in the history of the concept of "error" and its uses in statistics and quantitative psychology. These events include: the use of the error law in astronomy, the application of the error law to the study of human phenomena and the notion of "the average man," studies of variation in early psychophysics and physiological psychology, and advances in correlational methods.

2.1. The Astronomer's Error Law

The origins of mathematical statistics are often traced to the error law. This is akin to what is nowadays typically referred to as the "bell curve," "normal," or "Gaussian" distribution. In the 19th century, it came to be referred to as the "astronomical error law" as it was widely used by astronomers through incorporation into the method of least squares (Porter, 1986). In astronomy, observational errors were a subject of utmost importance, particularly with respect to the accurate estimation of the positions of planets and their orbits around the sun (Batten, 2015). As early as the 17th century, Galileo discussed the idea that there was a certain amount of "error" associated with the observations of planets (Read, 1985). In 1726, Gregory, an astronomer, described "error" as a deviation between some celestial object and an observer's representation (typically in mathematical terms) from that object (as cited in Denis, 2001). The issue of "error" was a central one for the field of astronomy throughout the 18th century, the aim of which was to obtain *accurate* representations of phenomena under study. For example, if one were to obtain repeated measurements of orbital elements, how would they determine which of the repeated observations was most accurate (i.e., which best reflected the true orbit of a planet)? If these observations were taken under identical circumstances, then one simply could take an average of the observations. However, because observations are obtained under varying circumstances, the issue of

determining the accuracy of a given measurement remains. In 1809, German mathematician Carl Friedrich Gauss argued that one solution to the problem was to identify the estimated orbit that minimized the sum of the squares of the residuals (the difference between the observed values of the orbital elements and the estimated ones; see Gauss, 1857). Prior to the 19th century, astronomers were quite hesitant to combine observations of planets (Stigler, 1986). The idea that examination of combined observations might improve accuracy was not accepted. Many astronomers believed that if observations obtained under varying circumstances were to be combined, the errors associated with those observations would also multiply and become exacerbated (Stigler, 1986). Thus, the introduction of the method of least squares in areas such as astronomy in the early 19th century was pivotal to the acceptance of combined observations in scientific practice.

The method of least squares was first introduced in an 1805[1] publication by mathematician Adrien Marie Legendre in which the method is described as a tool for "deduc[ing] the most accurate possible results from observational measurements" (as cited in Stigler, 1986, p. 13). Legendre proposed the following linear equation:

$$E = a + bx + cy + fz + \&c.,$$

Here, "E" refers to the amount of "error" tied to a single observational unit. Thus, the number of "error" equations would be equal to the number of observational units. "In which *a, b, c, f, &c.* are known coefficients, varying from one equation to the other, and x, y, z, &c. are unknown quantities to be determined by the condition that each value of E is reduced either to zero, or to a very small quantity" (as cited in Stigler, 1986, p.13). Based on this, Legendre proposed a method of solving for the unknowns by minimizing the sum of the squares of the errors. This method, he described, was useful if the number of unknowns was equal to the number of equations. In so doing, "a kind of equilibrium is established among the errors which, since it prevents the extremes from dominating, is appropriate for revealing the state of the system which most nearly approaches the truth" (as cited in Stigler, 1986, p.13). By 1815, the method of least squares was a common statistical tool used in the field of astronomy (Stigler, 1986).

[1] Gauss (1857) argued that he had been using the method of least squares as early as 1795.

Thus, in the early 19th century, "error" referred to the deviations of observed measurements from their true values (i.e., literally "the truth" as Legendre (1805, p. 73) described it). The "true" value was taken to represent the *accurate* value of the measurement. The error law, then, derives from the specification of the theoretical distribution that Gauss associated with such observational errors. Under this law, infinite and repeated observations (or measurements) would produce errors that took on a bell-curved, or "normal," shaped distribution, with the mean of the distribution equaling the true value (Stigler, 1986). The error law, or the normal distribution, is given by the exponential function: $\frac{1}{\sigma\sqrt{2\pi}}(e)^{\frac{(X-\mu)^2}{2\sigma^2}}$. This function had been developed in 1733 by De Moivre (see Porter, 1986; Stigler, 1986). De Moivre was a particularly pivotal figure in showing that increasing the number of observations (or measurements) would decrease uncertainties pertaining to those measurements. In other words, De Moivre showed that the variability associated with a normal distribution (in terms of distances of observed values from the mean) was a function of the square root of the sample size ($\sqrt{n}$; Stigler, 1986). As Stigler (1986) describes it, De Movire quantified uncertainty (where certainty is equivalent to accuracy) by proposing that the accuracy associated with measurements (i.e., the degree to which those measurements reflect the "true" value) was given by $\sqrt{n}$.

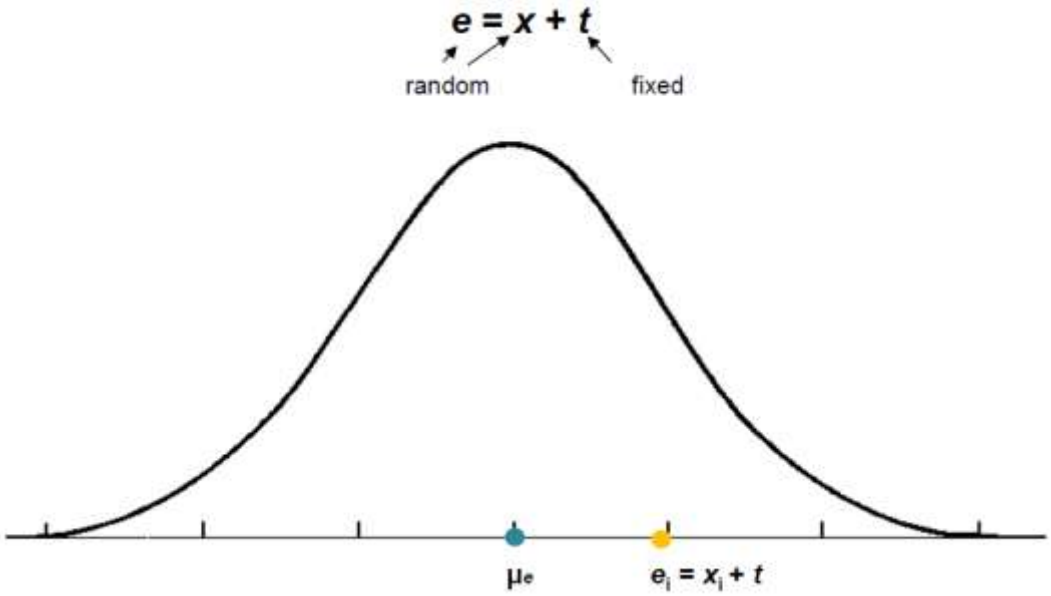
De Moivre's work was later expanded upon conceptually by Simpson (1755) who shifted focus from a distribution of observations to a distribution of errors, where the primary concern was the examination of the mean *error*. Thus, Simpson's (1755) important contribution to the conceptualization of "error" was that he conceived of it as a random variable with a given probability distribution.[2] According to Stigler (1986), this conception of "error" as a random variable allowed for "inverse inference" (p.101). That is, it allowed researchers to use probability to reason from an observed measurement what the "true" value might be. In mathematical terms, it meant that researchers could conceptualize the random variable of observed measurements, *O*, as the sum of a fixed true value, *t*, and a variable of errors, *E.* Thus, $O = t + E$, which then of course implies that $t = O - E$.

During the late 1700s, several different theoretical distributions were proposed by probability theorists to represent the error curve (the probability distribution associated

[2] Stigler (1986) notes that 3 different authors independently conceived of "error" as a random variable: "Simpson in 1755, Lambert in 1760, and Lagrange in about 1769" (p. 100).

with the error variable). Of particular importance was the work of French probability theorist, Pierre-Simon Laplace (Porter, 1986; Stigler, 1986). Laplace made several attempts at specifying an error curve, but was ultimately dissatisfied with his work and turned his attention to the central limit theorem in 1786 (Stigler, 1986). Then, in 1809, Gauss used De Moivre's (1733) exponential function for the normal distribution to represent the distribution of the error random variable (see Gauss, 1857). Laplace was inspired by Gauss' use of the normal distribution and came back to his work on the error curve during this time (Stigler, 1986). In particular, he used the central limit theorem to advance Gauss' idea of incorporating the normal distribution as the error curve within the method of least squares. Specifically, Laplace explained in a memoir, "if the errors of Gauss' formulation were themselves aggregates, then the limit theorem implied they should be approximately" normally distributed (as cited in Stigler, 1986, p. 143). Gauss, then, used the already familiar error law in the context of the method of least squares, which would lead to a surge in its popularity and use, especially in the field of astronomy (Hald, 2007). By the 1830s, it was widely accepted that the normal distribution curve was applicable to the distribution of errors of calculated averages *and* the distribution of errors associated with individual measurements (Porter, 1986). Given these advancements in early statistics, Figure 1 presents the distribution of the error random variable under the error law. Here, uncertainty is quantified as the standard deviation of the error random variable (i.e., the average amount of error).

Figure 1. Uncertainty Quantified: The Error Random Variable (The "Error Law")



Note. e = Error random variable; x = Observed score random variable; t = true score; μ_e = Mean of the error random variable; e_i = Error associated with observation i; x_i = Observed value of observation i.

Of importance for the proposed project is the way in which errors were conceptualized during this period. Laplace is often described as a Neo-Newtonian and determinist (Gigerenzer, 1987; Porter, 1986; Stigler, 1986). In his work, he described a philosophy of probability that implied a binary view of true values and errors; that is, true values were a product of permanent forces and errors were a product of accidental forces (Gigerenzer, 1987). True values existed independent of the observer and were yielded of permanent forces; errors were accidents that came about through the imperfections of human scientific practice. The purpose of mathematical statistics was to reduce such imperfections in measurement practice (Porter, 1986). The view that errors of observation were due to human ignorance was indeed a common view in the 18th and 19th centuries. Prior to the 19th century, however, errors of observation (and therefore human ignorance) were primarily attributed to the imperfection of measurement instruments. In the 19th century, we see the rise of the notion that errors could also come about from imperfections in the characteristics of the *observer*. This realization first came about when an astronomer named Maskelyne noticed that his assistant "observed the times of stellar transits almost a second later than he did" (Boring, 1929, p. 133). Although this difference in time was initially dismissed as inadequacies in the level of

expertise of the research assistant, Bessel later observed similar differences in observations even when observers were highly regarded experts in their field (Boring, 1929). Thus, differences in measurements obtained through human observation came to be known as the "personal equation" (Boring, 1929; Nunnally, 1970; Stigler, 1986). The implication for the concept of "error" was that human ignorance about objects of measurement could be seen to be due to many different sources, beyond imperfections with the measurement instrument itself. In 1838, Bessel identified 11 sources of random errors in the context of astronomical observations (as cited in Porter, 1986). These included instrumental flaws pertaining to telescopes as well as errors on the part of the researcher. Thus, observational errors in scientific practice in the 18th, and early 19th centuries were conceived of as being due to inadequacies in measurement practices, including imperfections of instruments as well as inadequacies of human observers (Porter, 1986).

2.2. Quetelet and L'Homme Moyen

The error law was subject to several important interpretations throughout the 19th century. As mentioned above, errors were only meaningful to astronomers and probability theorists in the 18th century insofar as they represented perceptual imperfections in humans that needed to be eliminated. In the 19th century, however, we see several gradual shifts in the interpretation of deviations from the mean. The first of these shifts is often attributed to the work of astronomer and mathematician Adolphe Quetelet. This section will detail Quetelet's uses of the error law and the influences his work had on the field of social statistics, the first area in which statistical tools were applied to the study of social and human phenomena.

Quetelet's (1842) work in social statistics was conducted primarily at the group or aggregate level (i.e., on a group of persons). He believed that the sciences should investigate collective objects rather than single objects. Quetelet (1842) viewed these collective objects as comprising a single social entity, and, he was particularly interested in studying the collective object of "man." He believed that society was a "social entity endowed with properties and tendencies that would not be significantly altered by political leaders" (Porter, 1986. p. 105). This implied that individual differences in human phenomena were irrelevant to the consistent character of collective society. He supported this idea by presenting demographic evidence of the unchanging character of

society year-to-year. Specifically, he used the error law to show that individual differences in human phenomena were normally distributed around an average value (Heidelberger, 1987). The consistency of birth and death rates over time, for example, was used as evidence that statistical laws operate at the collective level, regardless of differences at the individual level (Quetelet, 1842). Quetelet's efforts in applying statistical tools to social studies culminated into a book first published in 1835 titled, "Sur l'homme et le développement de ses facultés, ou essai de physique sociale" (the 1842 translation is titled "A treatise on man and the development of his faculties"). With this book, Quetelet (1842) aimed to build a foundation for the area of social statistics.

Quetelet (1842) gave special attention to the mean, or what he described as "l'homme moyen" (the average man).[3] He explained, "if an individual at any given epoch of society possessed all the qualities of the average man, he would represent all that is great, good, or beautiful" (Quetelet, 1842, p. 100). Thus, Quetelet described the mean as an ideal representation of society, and distinguished between the true mean and the arithmetic mean, where the former is the "true" average value of a distribution that follows the law of errors, and the latter is merely a calculation of the mean based on any arbitrary set of observations (Boring, 1929; Porter, 1986). Quetelet (1842) was uninterested in describing or making inferences about individual cases. Such cases were unimportant, he argued, as all deviations from the mean, including inter-individual differences, would inevitably cancel one another out. Instead, Quetelet (1842) believed that mediocrity was the ideal, and if one could examine a large enough group of individuals they would always obtain a normal distribution curve, thereby confirming the error law as a statistical law that governed the activities of humans.

Importantly, Quetelet applied the error law to *inter-individual* variation (i.e., observations of the same measurement instrument obtained from *different* objects or people). Moreover, he not only believed that normal distributions governed such inter-individual variation, but that the variation itself represented inaccuracies (i.e., "error" in the same sense as in astronomy; Porter, 1986). Thus, Quetelet believed that differences in specific cases could be ignored in lieu of the average value (Heidelberger, 1987). In other words, "we might regard such human variation as if it occurred when nature aimed

[3] The "average man" can actually be thought of as "average men" and "average women." Although Quetelet (1835) used the singular label "l'homme moyen," his work implied that there were a number of different "averages" depending on the faculties being examined (Stigler, 1986).

at an ideal and missed by varying amounts" (Boring, 1929, p.467-468). Quetelet (1842) carried forward the same interpretation of deviations from the mean arising from repeated observations of the same object to deviations from the mean arising from observations of different objects. As such, individual differences in height and criminal activity, for example, were merely interpreted as being imperfections of some ideal state of human nature.

Quetelet's contribution to statistics and to the social sciences was not a minor one. He was the first to show that statistical methods could be applied to social studies (at the aggregate level) in the same way that they were applied in the physical sciences. Quetelet firmly believed that a single procedure should be used in all fields of study, and he was convinced that statistics would serve as a unifying approach (Porter, 1986). Thus, Quetelet broadened the scope of statistical tools of analysis by applying the error law to the study of human and social phenomena, particularly at the level of inter-individual variation. That is, he argued that if researchers were to obtain a large enough number of values on a characteristic (e.g., height or weight) from a group of persons, the error law dictated that those values would form a normal distribution around the average (Quetelet, 1842). Quetelet went even further to propose that if values collected on a variable formed a normal distribution curve, this could be taken as indication that the variation observed was due only to accidental (independent and random) causes and therefore the observations were homogeneous (Stigler, 1986). Although Quetelet's work was highly influential in pushing forward the field of social statistics, his ideas regarding the applicability of the normal distribution were later debunked and met with harsh scrutiny. For example, in a 1922 paper, Edgeworth remarked,

> The theory [of errors] is to be distinguished from the doctrine, the false doctrine, that generally, wherever there is a curve with single apex representing a group of statistics – one axis denoting size, the other axis frequency – that the curve must be of the "normal" species. The doctrine has been nicknamed "Quetelismus," on the ground that Quetelet exaggerated the prevalence of the normal law (as cited in Stigler, 1986, p. 203).

Despite the criticism targeted at Quetelet's ideas regarding the average man and his uses of the normal curve, the empirical results of Quetelet's analyses were interesting in their own right. Indeed, it was the results of his analyses rather than his philosophical positioning that would come to influence the monumental work of British

biologist and statistician, Sir Francis Galton (Boring, 1929). I discuss Quetelet's influence on Galton, and Galton's contributions to psychology and the use of error, in a following section.

2.3. Fechner, Wundt, and the Study of Intra-Individual Variations

Quetelet may be recognized as the first to apply statistical tools to the study of human and social phenomena; however, Gustav Fechner in Germany is generally cited as the first to study *psychological* (primarily sensory) phenomena using measurement and statistics (see Fechner, 1860). Moreover, Willhelm Wundt, who was highly influenced by Fechner's work, is regarded as the first to have established a recognized laboratory of experimental psychology. The early works of Fechner and Wundt are often referred to jointly as the "Fechner-Wundt" tradition of psychology. Such early German psychology is perhaps most appropriately referred to as a form of "physiological psychology" given its emphasis on physiological and sensory phenomena (alternatively, Fechner (1860) referred to his own work as "psychophysics").

The Fechner-Wundt tradition, although often tied to the beginnings of psychology, is not entirely independent from earlier works in the physical sciences, particularly those of observational astronomy. For example, Fechner's work was influenced by the previously discussed "personal equation" phenomenon that arose out of astronomy (Boring, 1929). Given that the personal equation was a distance between the observed recorded measurements obtained by two different researchers, it had direct implications for reaction time research (i.e., observers differed in the amount of time it took them to react to stimuli). The rise of psychophysics and physiological psychology was highly influenced by the notion that differences in reaction time were due to underlying physiological and psychological characteristics (e.g., attention, expectation, preparation; Boring, 1929). Studies of the personal equation showed that researchers could quantitatively examine such underlying phenomena by using planned experimental methods (Stigler, 1986). For example, a researcher might manipulate the volume of a tone and ask participants to indicate when they noticed a change in the volume (Fechner, 1860). Interestingly then, it might be argued that one impetus for the ascendance of physiological psychology was the study of human phenomena that gave rise to errors in astronomical observation.

Under the Fechner-Wundt tradition of experimental psychology, the object of study was typically a single individual (i.e., *N*=1). For example, Wundt was interested in single-case studies as a means of elucidating the causal processes functioning within an individual's mind (Boring, 1929; Danziger, 1987). In addition, both Fechner and Wundt used factorial design and experimental methods to "control" for constant causes (Boring, 1929). This contrasts with Quetelet's method of using the error law to determine whether a distribution of scores was homogenous (and therefore not influenced by a constant cause; Stigler, 1986). In fact, Fechner was a strong advocate of modelling psychophysical phenomena with generalized, non-normal, distributions (Heidelberger, 1987). Moreover, unlike Quetelet, who mainly studied human phenomena at the *inter-individual* level, "error" in this context in which Fechner conceived of it referred to *intra-individual* variability.

It is in this early work in German experimental psychology/psychophysics that we first see the concept of "error" applied to the study of psychological and psychophysical phenomena. However, we also see a shift in the interpretation of the meanings attributed to such error. In much of the 19th century, "error" was conceived of as arising from a chance process that was due primarily to accidental forces outside of human control (e.g., ignorance of the observer or imperfections in measurement tools). It can be argued that such a view is in line with the deterministic philosophy that ruled the era of what has often been referred to as "mechanistic physics" (see Kruger, Daston, & Heidelberger, 1987). Under this view, indeterminism is only allowed insofar as it is "a result of human ignorance" (i.e., "error"; Heidelberger, 1987, p. 135). For example, under Quetelet's conception, society was viewed as an entity that was governed by constant causal forces in the same way that the orbits of planets were governed by constant gravitational forces (Heidelberger, 1987). Thus, "error" was conceived of as coming about from "accidental" and variable causes that resulted in perturbations in observed phenomena. Although such accidental causes were due to "chance" processes, they were interpreted as being governed by the error law (Porter, 1986).

Fechner's philosophical outlook has been contrasted with the deterministic philosophy of mechanistic physics. Heidelberger (1987) described Fechner as an indeterminist who believed that regularities in inter-individual differences in human phenomena (e.g., height and crime rate) did not necessarily indicate a lack of human freedom. Fechner believed that humans, and nature in general, were not bound entirely

by determined causes and that people had the ability to act of their own free will (Heidelberger, 1987). Fechner's notion of indeterminism "led to the conception that probability theory is an empirical science of chance phenomena in nature" (Heidelberger, 1987, p. 135). An excerpt from an English translation from Fechner's book "Kollektivmasslehre" ("Measurement of collective objects"; 1897, p.64) illustrates this view:

> A chance variation of the specimens, as I see it, is independent of any arbitrariness that arises in the measuring process as well as of any law of nature that interferes with the state of the values. It does not matter which of these variations plays a role in the determination of the objects; only variations that are independent of these changes happen by chance (as cited in Heidelberger, 1987, p.142).

Under Fechner's view, we see a shift in the interpretation of "chance", which in turn, influenced the interpretation of "errors." Like Quetelet, Fechner was interested in collective objects; however, unlike Quetelet, he interpreted variations from the "true" mean as potentially being due to chance that was separate from the imperfections of instruments or other sources of "random" error. Although much of the empirical work conducted under the Fechner-Wundt tradition was based on *N*=1 data, both inter-individual and intra-individual error was conceptualized as deviations from true values that were potentially characterizations of real natural processes with associated probabilities (Gigerenzer, 1987).

2.4. Galton, Pearson, and the Study of Inter-Individual Variations

I now turn to the tradition of research tied primarily to the works of Francis Galton in England during the 19th century. As mentioned previously, Galton was quite familiar with Quetelet's work with the error law and its use to study inter-individual variations. He was interested in error analysis; however, he did not believe that the error law dictated an idealistic "true" value at the center of error deviations (Porter, 1986). Instead, Galton was interested in error insofar as it represented variation amongst observations, and, predominately (if not solely) interindividual variations (Boring, 1929; Porter, 1986). According to Galton (1869), such variation was truly a meaningful product of nature in that it reflected variation in trait expressions and not merely random error or constant causes. Moreover, being highly influenced by the work of his cousin, Charles Darwin,

Galton viewed inter-individual differences not as “error” but as a sign of human progress. That is, the fact that variation could be seen amongst human phenomena (at the population level) was an indicator of human evolution and progression (i.e., natural selection requires variation amongst traits). As such, Galton’s (1869) aim was not to eliminate “errors” but to study them. He wanted to describe variation rather than limit his investigations to mean values (Galton, 1869; 1875; 1888). Thus, Galton can be considered the first to make use of error analysis to study variation (e.g., in his studies on heredity). He was also the first to use statistical tools such as correlation analysis in combination with questionnaires and mental tests, to study mental capacities (Danziger, 1987; Porter, 1986). Such correlational methods were later advanced mathematically by statistician Karl Pearson. However, readers should note that although development in correlational analysis is often credited to the works of Galton and Pearson, the mathematical foundations for correlation and regression analysis can be traced as far back as to Legendre’s previously described work on the method of least squares in 1805 (see Denis, 2001). Moreover, Pearson (1896) acknowledged that astronomer and physicist August Bravais had laid out a mathematical calculus for correlational analysis and described the “regression line” in 1846. According to Denis (2001), it would be most appropriate to state that, “Galton found empirically what Bravais deduced mathematically” (p. 43).

Nonetheless, the tradition of early British psychology is often referred to as the tradition of “correlational psychology” because of the methods that were used. Given that statistician Pearson was a pivotal figure in developing the mathematics of these methods, it is also often referred to as the “Galton-Pearson” tradition of psychology (Danziger, 1987). Under this tradition, inter-individual variability was the focus of research. Importantly, Galton and Pearson were not concerned with measurement theory. Pearson, for example, was a positivist who did not believe that the role of science should be to seek out unobservable values such as the “true score” (Gigerenzer, 1987). Instead, the Galton-Pearson tradition focused mainly on describing variations in human phenomena across individuals. The aim was not to provide evidence that such variation was driven by hidden and underlying causes; rather, the aim was to establish relationships between observed phenomena (Galton, 1869;1888). This being said, Galton (1869) wrote extensively on how people could be classified and differentiated in terms of their natural abilities, reputations, and race. Thus, his work indeed consisted of

many “causal” undertones in relation to factors such as intelligence and physical traits. Moreover, Galton (1869) explicitly used measurement and quantification as a means of claiming the superiority of certain individuals over others (e.g., based on race). As we will see, this is a feature that undoubtedly persisted into the 20th century with the proliferation of mental and intelligence testing.

Not only was Galton interested in examining variation in univariate distributions, he was perhaps even more interested in examining variation in bivariate distributions. Galton (1875, 1888) wrote considerably on methods for examining the distributions of *joint* errors. Prior to Galton (1888), others had examined the distributions of joint errors; however, none had interpreted the product of deviations on two random variables as an estimate of “co-relations” (Traub, 1997; Walker, 1929). Galton (1888), with the help of Pearson who would later advance the mathematical statistics (see Pearson, 1895) involved in correlation analysis, introduced the concept of correlation to mathematical statistics. Moreover, Pearson (1896) developed the following correlation coefficient, *r,* as an estimate of the degree to which two variables are linearly related.

$$r = \frac{S_{xy}}{\sqrt{S_x^2 \cdot S_y^2}}$$

Here, x and y represent deviations of values from their respective median values on two observed variables, X and Y. Thus, xy refers to the products of corresponding deviations (e.g., for each individual who is observed) on characteristics X and Y and S_{xy} refers to the sum of these products. In the denominator, S_x^2 and S_y^2 refer to the sums of the squared values of each of x and y. This notion of correlation, and Pearson’s *r*, would prove to be a pivotal concept for the burgeoning field of psychometrics.

In examining Galton’s uses of the error distribution in his studies on variation, one can clearly see differences in the conceptualization of “error” under the Galton-Pearson and Fechner-Wundt traditions. First, Fechner and Wundt were primarily interested in error distributions that were brought about from repeated measurements of the same event (intra-individual variability), whereas Galton and Pearson focused on error distributions of characteristics measured in groups of individuals (inter-individual variability). Moreover, Fechner and Wundt used error analysis as a “calculus of error,” whereas, under Galton and Pearson, error analysis was used as a “calculus of

exploration" (Danziger, 1987, p. 39). Despite their differences, the methodological traditions of the German and British studies of psychological phenomena in the 19th century both had profound impacts on 20th century psychology. In the following section, I trace the emergence of psychological test theory in 20th century America, explicating the influences of the German and British treatments of "error."

Chapter 3.

Error in Historical Perspective Part II: Twentieth Century Psychology and the Emergence of Test Theory

It was in the late 19th century that psychology began to establish itself as a discipline. By the early 20th century, psychology had begun to flourish in the United States. During this time, a new-found interest in the study of psychological or "mental" phenomena migrated from Great Britain and swept across North America. This interest can be explained by the socio-cultural and political climate of the time. Government officials, as well as the public, were interested in research that could be applied to the flourishing industries of education and military. As such, psychology took on a pragmatic role where the aim was to provide scientific findings that could be applied to growth industries. Fechner's factorial methods and Wundt's laboratory of experimental psychology were useful insofar as they provided a model for what scientific psychological research might look like. However, it was the correlational methods of Galton and Pearson that promised American psychologists practical results (Danziger, 1987; Gigerenzer, 1987). As Boring (1929) aptly stated, American psychology "inherited its physical body from German experimentalism, but it got its mind from Darwin" (p. 494). Nowhere did this influence have greater impact than on the field of psychological test theory (i.e., psychometrics) that would arise in conjunction with the proliferation of mental tests. Here, I trace the concept of error in 20th century psychology, focusing primarily on three general frameworks for thinking about measurement and error in psychology: classical test theory, item response theory, and generalizability theory. I also consider the influence of statistician Ronald Fisher's analysis of variance on the discipline of psychology.

3.1. Foundational Concepts in Classical Test Theory

Traub (1997) described the rise of classical test theory in 20th century psychology as being due to the culmination of three important prior events: (1) the acknowledgement of measurement errors; (2) the formulation of error as a random variable; (3) the

development of correlational methods. Indeed, we have seen how the first two events had occurred in the 18th and 19th centuries with developments in the use of the error law. Moreover, in the late 19th century, Galton conceptualized the notion of correlation, which Pearson later advanced in his mathematical treatments. Charles Spearman would then use such correlational techniques, in the context of his studies on "intelligence," to contribute to the development of classical test theory in the early 20th century. In this first sub-section, I introduce basic and fundamental concepts under the classical test theory framework.

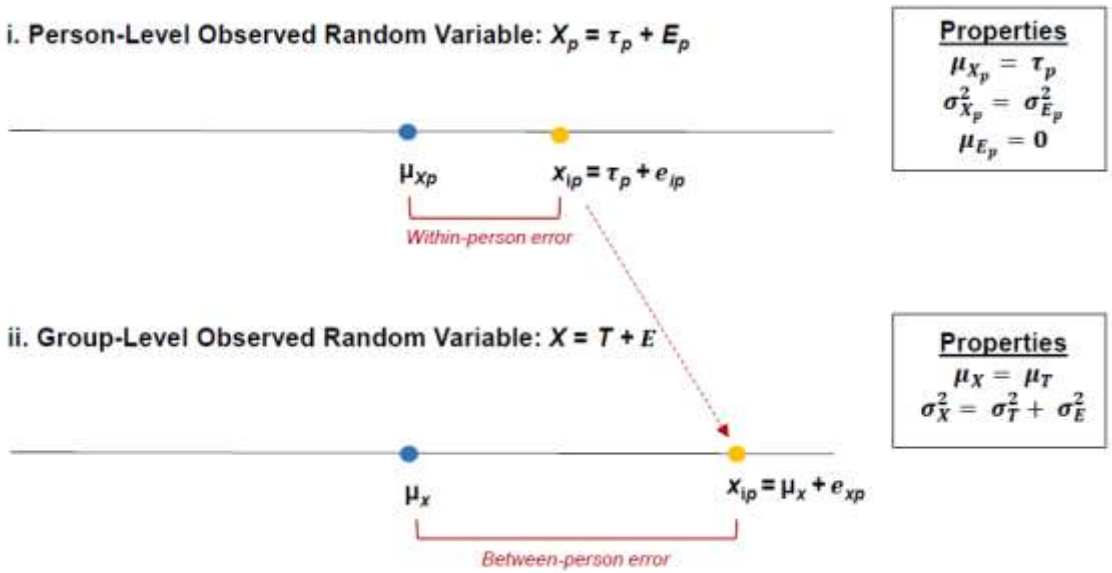
3.1.1. Error at the Person-Level (Intra-Individual)

At the person-level, the classical framework theorizes a propensity distribution of scores that is hypothetically based on infinite and repeated intra-individual observations (i.e., repeated observations obtained from the same person). Importantly, these observations are theorized as being obtained "in vacuo," such that extraneous factors, such as time, and systematic errors are irrelevant. Any source of variation is therefore conceptualized as being due to random error. Fundamental to classical test theory is the notion that person-level observed scores are composed of two parts. For example, a single person's observed score on a personality battery is composed of that person's true score and error of measurement. This proposition is mathematically formulated as: $x_p = \tau_p + e_p$ where x_p denotes an observed score for person *p*; τ_p denotes the true score for person *p*; and e_p denotes error of measurement pertaining to person *p*'s observed score (Gulliksen, 1950; Thurstone, 1932). Thus, X_p is a random variable composed of observed scores for person *p,* τ_p is a fixed constant, and E_p is also a random variable composed of errors of measurement for person *p.*

Several important properties of these concepts can also be identified. First, given that all values on the person-level random variable X_p cannot be directly observed in vacuo, a general shape for scores associated with this variable cannot be specified for each person, *p* (Lord & Novick, 1968; Sijtsma, 2009). Instead, it is merely assumed that the variance of this random variable, $\sigma^2_{X_p}$ is equal to the variance of the person-specific error random variable, $\sigma^2_{E_p}$. However, the means or expected values of these two distributions are not presumed to be equal. Rather, the mean of the person-specific observed score random variable, μ_{X_p} is equal to the true score for person *p*, τ_p and the

mean of the person-specific error random variable, μ_{E_p}, is equal to zero. Figure 2 presents the conceptualization of the distribution of X_p and related properties.

Figure 2. The Distributions of the Person- and Group- Level Random Variables X_p and X



Notes. τ_p = Person-specific true score; E_p = Person-specific error random variable; x_{ip}= A single observed score for person *p* on observation *i;* e_{ip} = Person-level error associated with observation *i*; μ_{Xp} = Mean of the person-specific observed score random variable; $\sigma^2_{X_p}$ = Variance of the person-specific observed score random variable; $\sigma^2_{E_p}$ = Variance of the person-specific error random variable. μ_{Ep} = Mean of the person-specific error random variable; μ_X = Mean of the group-level observed score variable; e_{xp} = Error associated with a single observation for person *p* at the group-level; *T* = True score random variable; *E* = Group-level error random variable; μ_T = Mean of the true score random variable; σ^2_X = Variance of the group-level observed score random variable; σ^2_E = Variance of the group-level error random variable.

3.1.2. Error at the Group-Level (Inter-Individual)

The components of classical test theory at the person-level (i.e., when repeated measurements are obtained on the same person) described above is based on theoretical assumptions about how person-level scores would behave if one were hypothetically able to obtain measurements in vacuo. Since this would be an impossible task to carry out, these assumptions at the person-level have no empirical basis and researchers are unable to obtain estimates of true scores in this way. In practice, researchers instead work at the group-level, i.e., they obtain single or composite measurements from *different* people. Observed scores obtained from multiple people consist of both variation from each of the persons' true scores (this is taken to reflect *real*

differences with respect to the quality measured) and variation due to errors of measurement pertaining to each person's true score (this is taken to reflect *random error* around the "true" measurement*;* Traub, 1994).

Thus, we can now consider the properties of the random variable, *X*, which is a function of two other random variables: $X = T + E$. Here, X is a variable of observed measurements at the group level. It is composed of *T*, which represents true scores of individuals on the given measurement, and *E,* which represents the difference between observed scores, *X*, and true scores, *T* (i.e., "errors" of measurement). Recall that at the person-level, τ_p is fixed; therefore, at the group level, *T* is a random variable based solely on differences *among* persons in their true scores. However, since E_p is random at the person-level, the randomness of E at the group level is based on both *within*-person and *between*-person variability. This means that the variance of the distribution of observed scores will be larger than the variance of the distribution of true scores because it is based on 2 sources of variation rather than a single source.

Finally, there are several important properties at the group-level that should be considered. First, the mean of the observed-score group-level random variable, μ_X, is equal to the mean of the true-score random variable, μ_T. Second, it is a basic assumption of classical test theory that errors of measurement should be uncorrelated with the true scores. This further implies that the covariance of *T* and *E* will be equal to 0. Thus, we can conceptualize the total variability in observed measurements, or the variance of the observed-score random variable *X,* as being equal to the sum of the variances of the true score random variable, *T,* and the error random variable, *E.* Mathematically, $\sigma_X^2 = \sigma_T^2 + \sigma_E^2$. Figure 2 presents the conceptualization of the distribution of *X* and related properties.

3.1.3. Precision, standard error, and reliability.

The treatment of "error" under classical test theory gains greater meaning with the consideration of the concepts of precision, standard error, and reliability. Under classical test theory, the standard error of measurement refers to the variability associated with the person-level error random variable, E_p (Traub, 1994). In other words, it is the standard deviation of the person-level error variable which is equal to the standard deviation of the person-level observed scores variable. It gives us a sense of

the size, on average, of the errors associated with the person-specific observed variable, X_p. This concept is not an entirely new one. The notion that the amount of "uncertainty" associated with observations on a random variable of interest can be captured by referring to the variability of an error random variable was introduced early in the history of mathematical statistics (Stigler, 1986). Classical test theory takes this idea a step further by tying the concept of "precision" to it. Precision can be conceptualized as referring to how close a person's observed values are to their true or expected value. Less variability (i.e., smaller standard error) is an indication of greater precision.

It is important to note that, under the classical test theory framework, the person-level propensity distribution is not actually modeled in practice; rather, it serves as a conceptual starting point that is later formalized at the group level in terms of reliability. Thus, the above descriptions of standard error and precision are expanded to the group level. That is, the dispersion of scores is considered only at the group-level random variables, *X*, *T*, and *E*. In particular, the variability associated with the random variable, *E*, gives a sense of the size, on average, of the errors associated with the group-level observed variable, X. Reliability then, is computationally defined at the group-level as the ratio of true score variance (σ_T^2) to observed score variance (σ_X^2; Kelley, 1924; Thurstone, 1932; Traub, 1994). Given that group-level observed scores are made up of both true score variance and error variance, larger values of this ratio indicate smaller amounts of error and, thus, a more reliable estimate of true scores.

What exactly is meant by a "reliable" estimate? In a general sense, reliability can be thought of as referring to the degree to which the same intra-individual measurement is repeatable over hypothetically repeated measurements (Nunnally, 1970). More specifically, Lord and Novick (1968) described reliability as an indicator of how repeatable an *individual's* performance on a test is. The amount of repeatability is given by the individual-level propensity distribution (i.e. the distribution of X_p). Theoretically, repeated measurements will never replicate one another exactly. This implies that measurements will always have some amount of "unreliability" or "error" associated with them (Stanley, 1971). However, given that repeated measurements cannot be attained in vacuo at the individual level, like precision and standard error, reliability is a concept that is theorized at the individual-level, but is formalized in practice at the group-level.

Finally, some have argued that there are different "types" of reliability, based on the computational method used. Psychometricians distinguished between "coefficients of equivalence" and "coefficients of stability" in the early 20th century (Sijtsma, 2016). The former is calculated based on items from two different tests that are taken to be measuring the same thing; whereas, the latter is calculated based on scores obtained on items from the same test but at differing points in time. Cronbach (1951) famously extended these ideas by describing two further types of reliability: "coefficients of equivalence *and* stability" are those that are calculated based on two different sets of items from two different tests that are administered at two different time points; "coefficients of precision" refer to scores obtained from the same test at the same time point from the same people. However, given that it is arguably impossible in practice to administer the same test twice, at the same time, to the same person, this final coefficient is taken to be theoretical.

3.2. Error under Classical Test Theory

Thus far, I have outlined various concepts foundational to the logic and framework of classical test theory. In this second sub-section, I draw from these foundational concepts to discuss major theoretical advancements involving the concept of "error" that were made under classical test theory. I begin with a focus on contributions by Charles Spearman, discussing his classification of accidental versus systematic errors, as well as his contributions to reliability and what came to be known as the "Spearman-Brown prophecy." In the years following Spearman's initial work, numerous authors developed and published important works extending the concept of reliability, as well as test theory, both mathematically and practically (e.g., Abelson, 1911; Cronbach, 1951; Guttman, 1945; Kelley, 1921; Kuder & Richardson, 1937). I will focus on explicating the ways in which "error" and related concepts were described and conceptualized under test theory following the initial works by Spearman.

3.2.1. Charles Spearman: "Accidental" vs. "Systematic" Errors, Correction for Attenuation, and the "Reliability Coefficient"

As already described, by the early 20th century, the notion that measurements consist of some amount of "error" was generally accepted amongst scientific communities. Moreover, the notion of "co-relation" established by Galton and later

advanced by Pearson was also accepted as a legitimate methodological concept. This is evidenced by the fact that during this time, numerous authors were beginning to publish papers focusing specifically on the correlational method (Traub, 1997). Of particular importance to measurement in psychology were Spearman's (1904a; 1904b) works on intelligence and his correction for attenuation formula. Spearman (1904a) described how measurements of psychological abilities would vary if one were to take repeated measurements from the same person. He described these variations as being "accidental" in nature, making reference to one of two types of "error:" "systematic" and "accidental" (Spearman, 1904a, p. 88). Systematic errors are inaccuracies of measurement that are "constant" and explained by factors not measured by the test at hand; whereas, accidental errors are inaccuracies that are "variable" (i.e., differing between each individual case) and not necessarily explained by any external factors (Spearman, 1904a, p. 88). Spearman (1904a) argued that if one were to compute a correlation coefficient for independent measurements obtained on a group of people, the resulting coefficient would always contain some amount of accidental error. Moreover, such accidental errors have an attenuating effect on the "true" correlation, i.e., they make computed correlations between measurements obtained on two variables appear smaller than they really are. Thus, Spearman (1904a) developed his methods of correcting for attenuation in pursuit of obtaining the "real" correlation between two observed variables. These attenuation formulae were meant to correct for the accidental error components of two independent sets of observations, thereby producing a better estimate of the "real" correlation (Spearman, 1904a).

As an example, Spearman (1904a) stated, "suppose that we wish to ascertain the correspondence between a series of values, *p*, and another series, *q*" (p. 90). The correlation attenuation formula is then given as:

$$r_{pq} = \frac{r_{p'q'}}{\sqrt{r_{p'p'} \cdot r_{q'q'}}}$$

Here, $r_{p'q'}$ is the average correlation between the observed values on *p* and *q,* while $r_{p'p'}$ and $r_{q'q'}$ are "the average correlation[s] between one and another of these several independently obtained series of values" for p and q, respectively (Spearman, 1904a, p. 90). Spearman's initial conception of the correction for attenuation formula was met with harsh criticism from Pearson. In particular, Pearson (1904) criticized Spearman

for not providing a proof of his formula. In response, Spearman published another two papers in 1907 and 1910. The first, in 1907, consisted of a proof for his correction of attenuation formula; however, it received further criticism on the grounds that his correction did not account for errors that were not "accidental" in nature (e.g., Brown, 1910). In 1910, Spearman argued that Pearson had primarily been concerned with "the theory of sampling errors" but had "scarcely touched on the errors of observation" (Spearman, 1910, p. 283). He again provided his proof for the correction for attention formula and further emphasized the distinction between "accidental" and "regular" errors, clarifying that the correction was only suitable for dealing with errors of the "accidental" type (p. 273).

Spearman also introduced the term "reliability coefficient" in his 1910 paper. He defined this as "the coefficient between one half and the other half of several measurements of the same thing" (Spearman, 1910, p. 281). Spearman proposed dividing measurements obtained on the same thing from the same person into two groups. This division, he stated, should be "made in such a way, that any differences between the different group averages (for the same individual) may be regarded as quite 'accidental'" (Spearman, 1910, p. 274). Spearman (1910) further provided a formula for calculating a reliability coefficient as a function of the length of a given test. This result was also independently presented by Brown (1910) in a paper published in the same edition of the same journal (*The British Journal of Psychology*). Thus, the result came to be known as the "Spearman-Brown Prophecy." Over the next several decades, researchers extended the notion of reliability under the classical test theory framework, developing a variety of reliability coefficients (see Thurstone, 1932; Gulliksen, 1950 for an overview), much of this work focused on clarifying and extending "the meaning of the key phrases 'one half,' 'other half,' and 'same thing'" in regards to reliability (Stanley, 1971, p. 370).

Spearman's (1904a, 1904b, 1910) works are an illustration of how important the concept of "error" became for psychologists in the early 20th century. It allowed psychologists to deal with measurement issues by introducing new concepts such as "reliability." This was, of course, undertaken with the help of correlational methods formalized earlier by Galton and Pearson. However, as mentioned above, Pearson was not fond of Spearman's use of correlational methods (see Pearson, 1904) and his dislike of Spearman's work was perhaps fueled by the fact that Spearman was using

correlational techniques to examine unobservable psychological phenomena (i.e., intelligence). The use of correlational methods in such a pursuit unquestionably angered Pearson who held strong positivistic views.

Regardless, if psychology was to be considered a science in the early 20th century it would need to establish measurement practices. This would prove to be no easy feat. Spearman was not recording observations of planets; he was instead attempting to measure the historical, cultural, and *human* phenomenon of intelligence (although he likely did not view it as such). Like measurements in astronomy, it was presumed that there was a certain amount of "error" associated with measurements obtained from *people*. Indeed, under classical test theory, the mean of the person-level propensity distribution represents the expected value or "true score" for that individual and deviations from this value represent error. However, it is impossible to model the person-level propensity distribution in practice. That is, one cannot observe psychological behaviors and traits devoid of any temporal or environmental influence. It is also impossible to satisfy the assumption of independence of errors when working with repeated measurements taken from the same person. How then, could one judge the *accuracy* of measurements of complex human and psychological phenomena? Observations would have to be obtained from *different* people and "error" would need to be treated at the inter-individual level. Moreover, two different kinds of error would need to be identified, systematic and accidental, and a correction for the attenuation caused by accidental errors would need to be introduced. Spearman's contributions to the application of statistical "error" in psychology thus served as the foundational starting point for the flourishing field of psychometrics.

3.2.2. Truman Kelley: Defining "True" Scores

Although Spearman was a key figure in the development of classical test theory, he did not explicitly deal with the conceptualizations of statistical concepts such as "error" and "true" score for psychology. Recall that "true" scores had historically been conceptualized in varying ways. For example, the measurements obtained by an astronomer observing the positioning of a planet in relation to the sun would differ from the true, "accurate" position of the planet by some amount of error. If the astronomer were to obtain infinite measurements of the location of the planet, these repeated observed scores should be distributed around the true score (mean). Thus, the

conventional view in astronomy was that the "true" score was the *actual* value of the sought-after parameter as it *truly existed* in reality. In addition, this true score is reflected computationally by the average of repeated measurements of the *same* thing. Quetelet (1842) was influenced by this interpretation; however, he described the true score, or average value, as representing a hypothetical ideal. In addition, his studies were conducted at the group level, and therefore, *l'homme moyen* was based on the average of observations obtained from *different* people. For example, Quetelet (1842) believed that if one were to sample the heights of a large enough number of individual males, the average of those heights would represent the ideal height for men. This ideal height may or may not have actually existed in reality; the "true" value was simply a hypothetical ideal.

Although intriguing, Quetelet's conception of true score was not adopted into classical test theory. Kelley (1921) was perhaps the first to formally define true scores in the context of classical test theory. Kelley (1921) defined reliability as "the extent to which the test measures that which it in reality does measure – not necessarily that which it is claimed to measure" (p. 370). He further distinguished reliability from the concept of validity, stating that validity is concerned with whether "a test measures what it purports to measure"; whereas, reliability is concerned with "how accurately a test measures the thing which it does measure" (Kelley, 1927, p. 14). Thus, the true score is associated with what the test *actually* measures, regardless of whether that thing is the same or different than what a researcher *intends* for the test to measure. Kelley (1921) explained:

> The highest possible correlation which can be obtained… between a test and a second measure is with that which truly represents what the test actually measures – that is, the correlation between the test and true scores of individuals in just such tests. These true scores may be defined as the average scores of individuals upon a very large number (an infinite number) of just such tests (p. 372).

Both Kelley's (1921) definition and Spearman's (1904; 1910) descriptions of "true" or "real" correlations bear greater resemblance to the conception of true scores under early astronomy than to Quetelet's (1842) hypothetical ideal. That is, the "real" correlation is taken to *truly exist* independent of observer influence. True scores are taken to "truly represent what the test actually measures" (Kelley, 1921, p. 372). However, such "real" or "true" values are unobservable to the human observer because

they are obscured by some degree of measurement error (Thorndike, 1966). It is for this reason, perhaps, that Kelley's (1921) definition placed emphasis on the average of an infinite number of repeated measurements. That is, it emphasized the true score as arising from a hypothetical empirical process (the average computed over repeated and infinite measurements). Moreover, it is interesting and useful to note that Kelley, like other psychometricians at the time, felt the need to distinguish between what a test *purportedly* measures and what a test *truly* measures. This is a distinction that is unique to the field of psychology.

3.2.3. Louis Thurstone: Summarizing Ideas in Classical Test Theory

Indeed, reliability and validity became the cornerstone concepts of psychometrics. Thurstone's (1932) book, "The Reliability and Validity of Tests" was the first to take on the task of summarizing (and, in some cases, expanding upon) the major works produced under classical test theory in the early 20th century. This work provides a good sense of common ways that the concepts of "error," "true score," and "reliability" came to be viewed under this framework. Thurstone (1932) described the difference between "chance" or "random" errors and "systematic errors." He explained,

> Measurement in psychology is usually made with the handicap of unknown and uncontrolled factors so that the measurements are rather unstable… Since the unpredictable chance factors have a marked effect on psychological performance of all kinds it is necessary for us to pay even more attention to the magnitude of chance errors in psychological measurement than is necessary for many problems in the exact sciences (p.1).

Thus, "chance" errors, as Thurstone (1932) saw it, were especially critical for psychology. He further described the logic upon which the notion of reliability was founded, stating that unlike the physical sciences, one could not administer the same psychological test to the same person in hopes that differences in observed scores would only be due to chance variation. If a person were to take the same test twice, it is possible that they would do better on the second trial because of "increased familiarity with the nature of the examination" (Thurston, 1932, p. 2). One way to circumvent this problem is to administer two comparable tests and to interpret the difference in scores on the two tests as a measure of chance variation. However, this of course assumes that

the two tests are truly measures of the same thing and are, in a sense, identical to one another. Again, this is an assumption that is not easily justified in practice.

Thurstone's (1932) summary of classical test theory was important because it established disciplinary norms for thinking about concepts such as reliability and error in psychometrics. In doing so, it also helped to establish classical test theory as a legitimate way of thinking about measurements obtained from psychological tests. However, adopting the concept of "error" for the study of psychological phenomena was not an easy task and Thurstone's (1932) monograph clearly portrays some of the difficulties with attempting to obtain accuracy with psychological measurements. Dealing with uncertainty using quantitative analytic tools proved to be especially complex when the subject matter being considered is psychological in nature.

3.2.4. Louis Guttman: Error Based on Three Sources of Variation

Despite the challenges faced with quantifying psychological phenomena, psychometricians continued to discuss and develop ways of advancing classical test theory toward the latter half of the 20th century. In 1945, Guttman provided a clarification of the concept "error" for classical test theory through the identification of three sources of variation: trials, persons, and items. He described "error" as being defined in relation to each *person* for each *item* on each given *trial* in a population (or "universe") of trials. Guttman (1945) then explained that "unreliability" referred to the variation observed *across* trials. The implication for reliability was that one couldn't obtain a reliability coefficient solely based on measurements acquired from items under a single trial. However, since it is difficult in practice to obtain measurements from multiple independent trails, Guttman's (1945) aim was to determine what information *could* be obtained from a single trial. To this end, he formulated six different "lower bounds" to reliability.

Guttman's (1945) paper is important to the conceptualization of "error" under classical test theory for several reasons. First, he presents an attempt at distinguishing between different *levels* of analysis and therefore different kinds of error associated with each of those respective levels. Implicit in this discussion is the idea that "error" at different levels of analysis has different uses and therefore different meanings. Guttman's (1945) aim was to determine what researchers *can* establish about reliability

based on a single trial. However, the very fact that he emphasized the limitations of working with a single trial speaks to the inconsistencies between the logical foundations of classical test theory and actual research practice. That is, the logic of classical test theory is built up from the individual-level with consideration of repeated observations from the same person over multiple trials. Yet, in practice, researchers are typically working with observations from different individuals obtained during a single trial. This presents an issue for psychometricians because repeated observations cannot actually be obtained at the individual level in vacuo; thus, researchers must find ways to deal with individual-level measurement error while working at the group level. The introduction of various "lower bound" estimates of reliability by Guttman (1945) is representative of the ways in which psychometricians throughout the 20th century tried to handle such issues in classical test theory, i.e. by proposing different "kinds" of reliability estimates that could be used despite having only obtained observations from different people during a single trial. Notably, Guttman's (1945) distinction between different types of error operating at different levels of analysis is an important precursor to aspects of his later work in which he would highlight the consideration of sampling error for test theory. I discuss this work, and how it contributed to the development of generalizability theory, in a later sub-section.

3.2.5. Harold Gulliksen: Summarizing Ideas in Classical Test Theory

A second summary of advancements in classical test theory, following the initial publication by Thurstone in 1932, was published in 1950 by Harold Gulliksen. Gulliksen (1950) was heavily influenced by Thurstone's works and teachings. He attributed much of his work in "Theory of Mental Tests" to Thurstone, stating that Thurstone's contributions to classical test theory provide "confidence that psychology is beginning to take its place among the older sciences" (Gulliksen, 1950, p. ix).

Gulliksen (1950) begins his book by describing the aim of the psychometrician as being the determination of the "accuracy" of an observed score. Gulliksen (1950) goes on to reiterate the point made by Thurstone (1932) that psychological measurement contains much greater amounts of error than measurements obtained in the physical sciences. The "true" score, according to Gulliksen (1950) is "some number" (i.e., unknown and unobservable) that is taken to be an individual's "correct score" on a test (p. 5); "error," then, is the difference between what one observes (i.e., the score a

participant obtains on a test) and the true score. Moreover, like the "true" score, "error" is also unobservable; it is impossible to obtain repeated and infinite measurements at the individual level.

Gulliksen's (1950) descriptions of the concepts "error" and "true" score are reminiscent of the uses of these concepts under early statistics and observational astronomy. However, there is a clear attempt being made by Gulliksen (1950; and other psychometricians at the time) to mold these concepts to the study of psychological phenomena. Indeed, Gulliksen (1950) implies that dealing with psychological phenomena is unique in that observations obtained from psychological tests will typically contain greater amounts of error than observations obtained from instruments in the physical sciences. Nonetheless, this does not stop Gulliksen (1950) from using measurement and statistics in the pursuit of accuracy. From his perspective, psychology is establishing itself among other sciences by developing test theory.

While classical test theory was crucially important to the development of psychology in the 20th century, it did not completely resolve the problem of measurement for psychology (i.e., the problem that psychological measurements cannot be obtained in vacuo). In the second half of the 20th century, psychometricians began to argue that psychological measurement is much more complex than originally presumed by classical test theorists. As a result, alternative approaches to test development, notably item response theory (IRT) and generalizability theory (g theory)—which are described below—were advanced.

3.3. Meanwhile in Statistics: Fisher's Analysis of Variance

It is worthwhile to take a brief intermission from test theory to explore advances made in the field of statistics in the early 20th century that would become highly influential for the field of psychology. What I am especially referring to here is the method of Analysis of Variance (ANOVA) articulated by Ronald A. Fisher in a 1918 paper and later popularized through his seminal 1925 book, "Statistical Methods for Research Workers." This work is regarded by many modern-day statisticians as one of the most (if not *the* most) important pieces of literature for the field of statistics.

Fisher's (1925) book begins with an explanation that statistics is concerned with the study of populations, variation, and the reduction of data. He clarified what he means by populations by stating that "statistics is the study of aggregates of individuals, rather than of individuals" (p. 2). This is evidenced, he argues, by the fact that statistics is performed by obtaining *repeated* observations and by calculating aggregate-level statistics (e.g., means and standard errors). Clearly, then, statisticians are not concerned with individual results but with populations of all possible results. Given that statistics is performed on multiple observations, it is only natural that the field would be concerned with studying variation. Fisher (1925) further contrasts the goals of modern statistics with those of early statisticians, explaining that it is a feature of modern statistics to be interested in studying variation rather than merely obtaining averages. Here, Fisher (1925) is perhaps drawing comparisons between the works of Quetelet (old statistics) and Galton (new statistics). Indeed, Fisher (1925) references and draws upon Galton and Pearson's work in correlational methods throughout his book, indicating that his methods had been profoundly influenced by the correlational tradition.

Fisher (1925)'s book might be viewed by some as an introductory statistics textbook for substantive researchers. In fact, modern introduction to statistics textbooks in psychology are reminiscent of its structure. The book begins with basic concepts in statistics, diagrams, and distributions, and goes on to outline a variety of statistical tests such as chi-square, tests of differences between means, and the intraclass correlation. Most importantly, Fisher (1925) outlines the method of ANOVA and its applications for research. Under ANOVA, the total variability in an outcome is separated into parts and each of these parts is conceived of as being based on different causes, one of these being random error. This random error is conceptualized at the group level. For example, in the classic one-way ANOVA set-up there is one continuous outcome variable and one categorical predictor with k groups. The amount of variability in the outcome variable that is due to error is equal to the sum of squared distances, where the distance is between an individual's score on the outcome and the mean of the k^{th} group that the individual belongs to. Thus, Fisher's ANOVA is based on inter- rather than intra-individual variation. Under Fisher's model, the distance between an individual's score on an outcome and the mean of that individual's own hypothetical person-level propensity distribution on the outcome is not considered. That is, error on the person-level observed variable X_p is not considered. However, this source of variation should still conceivably

contribute to variations observed at the group level, i.e., theoretically, observations at the person level (if hypothetically measured in vacuo) would vary over repeated measurements. Thus, this intra-individual variation would still be captured to some extent in a single person's score, which would contribute to inter-individual variation at the group level.

Fisher's ANOVA was quickly adopted in the field of psychology and continues to be one of the most popular statistical tools used today. In many ways, ANOVA combined the best of both worlds when it came to the Fechner-Wundt and Galton-Pearson traditions (Danziger, 1987). It allowed for the successful integration of factorial design while also making use of tools for examining variation. Moreover, it was an accessible tool that could be easily adopted without extensive mathematical training or knowledge. Combined with the procedure of null hypothesis testing, it has become the go-to analytic technique for experimental psychologists.

3.4. Error under Item Response Theory (IRT)

In the latter half of the 20th century the use of ANOVA in experimental psychology was commonplace. At the same time, psychometricians continued to advance and develop test theory. In 1953, Guttman published a review of Gulliksen's (1950) "Theory of Mental Tests" in which he emphasized several major points of concern. Among these were criticisms that classical test theory had not provided a comprehensible or structured theory for the development of tests (Gulliksen (1950) also acknowledged this in his work) and that psychometricians had neglected the issues of sampling error and generalizability. The introductions of IRT and g theory in the latter half of the 20th century can be considered two ways in which psychological methodologists attempted to address these issues (IRT mainly tackled the issues of structure and item-level analysis, while g theory tackled the issues of sampling error and generalizability). In doing so, the concept of "error" was also advanced under these new frameworks.

It is important to note that whereas classical test theory is regarded as being an error model, many scholars would argue that IRT is not (Zumbo, personal communication). That is, the logic of classical test theory is strictly tied to the aim of reducing and, ultimately, eliminating errors. Although the concept of error is not abandoned under IRT, the focus is not on test-level error but rather on the performance

of items. Nonetheless, IRT extends the conceptualization of "error" by introducing the notion of "information." IRT (sometimes referred to under the umbrella term of "modern test theory") refers to several different item response models that deal with binary outcomes. More recently, it has also been extended to handle graded polytomous and multinomial responses. The beginnings of IRT can be seen in multiple works that emerged in the mid-20th century (e.g., Birnbaum, 1968; Birnbaum, 1969; Guttman, 1950; Lawley, 1943; Lazarsfeld, 1950; Lord, 1952; Lord & Novick, 1968). For example, in 1968, Lord and Novick published "Statistical Theories of Mental Test Scores" for which the main impetus was to formalize a coherent way of synthesizing the contributions of test theory. To this end, the book also contained contributions from Allan Birnbaum on his work on "latent trait theory." Birnbaum's (1968) writings were some of the first pieces of literature to lay the groundwork for IRT.

Of importance for the proposed project is the conceptualization of "precision" under IRT models. Lawley (1943) described how an individual's probability of correctly answering a dichotomous item is dependent on his or her level of ability in relation to a latent variable of interest as well as on characteristics of the item itself. Here, a "latent variable" is defined as an unobserved variable that is taken to have causal effects on a participant's performance on a given test (Lord & Novick, 1968). From an IRT perspective, precision of measurement is theorized to vary across the latent trait dimension. Thus, IRT does not presume a uniform value of precision for all test-takers.

Given that reliability is taken as an indicator of the amount of precision for a given test, the conceptualization of precision has implications for reliability. Recall that under classical test theory only a single index of reliability is calculated and this single index typically speaks to the *average* (i.e., across individuals) reliability of a test (i.e., it is based on the ratio of the aggregate-level true score variance and observed score variance). This index does not account for the fact that precision might vary based on an individual's level of ability on a desired trait or based on the characteristics of items on the test. IRT accounts for these factors by considering item information functions that specify the amount of "information" associated with a given item based on varying levels of ability on a latent trait (Birnbaum, 1968).

The concept of "information" was first described by R.A. Fisher (1922) as the reciprocal of the standard error of measurement. Under the IRT framework, information

is calculated based on specific levels of ability. Given that less standard error denotes greater precision associated with an estimate, the greater the information, the more precision (i.e., less amount of error) associated with measurements at a specified level of ability (Birnbaum, 1968p; Lazarsfeld, 1950). In other words, if the level of information associated with a given level of ability is large, it implies that measurements can be obtained with great precision for any individual who truly possesses that level of ability on the latent trait. If one were to plot the item information function (i.e., level of information against values of the continuous latent variable) they would find that more extreme values of the latent variable should have more error and less precision associated with them, whereas, values that are more common should have less error and greater precision (Slaney, 2006; Mellenbergh, 2011).

The use of "information" under the IRT framework thus extended the conceptualization of error in that it allowed for researchers to think of error as a value that varies across different levels of ability on a latent trait. In this sense, one might think of advancements to the "error" concept under IRT as another attempt at dealing with the complexities of "error" associated with unobservable psychological phenomena. Indeed, Lord & Novick (1968) explicitly stated that these are the *main* reasons for the existence of test theory.

> One reason we need to have a theory of mental testing is that mental test scores contain sizable errors of measurement. Another reason is that the abilities or traits that psychologists wish to study are usually not directly measurable; rather they must be studied indirectly, through measurements of other quantities (p. 13).

From this view, insofar as the phenomena of interest to psychological researchers is considered to be unobservable and incapable of being directly measured, the need for a coherent test theory that captures the complexities of error associated with psychological measurements will remain.

3.5. Error under Generalizability Theory

Generalizability theory (g theory) is an approach to reliability analysis developed by Cronbach, Gleser, Nanda, and Rajaratnam (1972) that is based on the work of Fisher (1925) in developing analysis of variance (ANOVA) methods (as well as others who later advanced ANOVA methodology, e.g., Burt, 1955; Cornfield & Tukey, 1956; Stanley,

1962). As previously described, under classical test theory, an observed score is theorized to be composed of two parts: a "true" score and random "error." Under this conceptualization, the error component of the observed score is "looked on as a sample from a single undifferentiated distribution" (Cronbach et al., 1972, p.1). Cronbach et al. (1972) argued that this notion of an undifferentiated "error" is too simple and ill defined. They further argued that what might be considered random "error" in one test analysis might be a source of interest in another. Based on these criticisms, Cronbach et al. (1972) advocated a Fisherian approach to test analysis under which "random" error could be modeled as attributable to a variety of sources. They proposed that researchers could "estimate how much variation arises from each controllable [error] source" (p. 1). This is done by adopting the ANOVA procedure of partitioning the total variability in an outcome into different components, which allows researchers to estimate the amount of variation in the outcome that is due to personal characteristics, different sources of error, and the interactions between each of these factors.

Importantly, generalizability theory also extends the conventional approach to sampling theory. Guttman (1953) had previously noted problems with the conventional view on sampling under classical test theory.

> Current sampling theory by itself cannot solve many problems of prediction and external validity. Conventional sampling problems concern the selection of people from a large population. Mental test theory faces also another type of sampling problem – that of selecting items from one or more indefinitely large universes of content. This is a basic problem of item analysis (p. 129).

To address the issue outlined by Guttman (1953), under generalizability theory, the term "population" is reserved to refer to *populations of subjects* and the term "universe" is used to refer to the "*universe under which the subjects might be observed"* (Cronbach et al., 1972, p. 9). In other words, there exists a universe of possible observations, and a "universe score" refers to a person's average score based on all possible and acceptable observations. An observed score, then, is composed of a person's universe score and multiple other sources of error. Observed scores are used to make inferences about universe scores; thus, under generalizability theory, "the question of 'reliability'… resolves into a question of accuracy of generalization, or generalizability" (Cronbach et al., 1972, p. 15).

In many ways, generalizability theory was an attempt to unify experimental and measurement work in psychology. As Cronbach et al. (1972) noted, ANOVA techniques gained popularity in experimental studies of psychology in the 20th century, and yet, studies of measurement maintained a classical test theory approach based on correlational methods. Part of the reason for this divide, explained Cronbach et al. (1972), was that experimental studies in psychology "regard subjects (persons) as a source of "error" in their analyses" (p. 2). In contrast, studies of measurement are "interested chiefly in the person tested and only secondarily in the conditions of observation" (Cronbach et al., 1972, p. 2).

3.5.1. The Problem of "True" Scores

The shift in thinking about the "true" score as a "universe" score under generalizability theory was implemented to emphasize that researchers make an inference about the population-level "universe" score based on the sample-level "observed" score. In addition, Cronbach et al. (1972) take into consideration the possibility that there may be more than one universe of scores to which an observed score might be generalized. However, it must be acknowledged that the question of what is implied by a "true" score is often a philosophical one. Lord and Novick (1968) described three different views on true scores within the psychometric literature. Thorndike (1964; 1966) argued that, due to their unobservable status, true scores are "mystical" and have no theoretical relevance to test development. Loevinger (1957), on the other hand, argued that the fact that true scores are unobservable implies that psychometricians need not concern themselves with them at all. Thus, the only kinds of questions that ought to be answered by psychometricians are those that concern that which is observable, i.e., observed scores. Whereas Thorndike (1964; 1966) did not find true scores to be theoretically relevant, Lord and Novick (1968) interpreted Loevinger (1957) as implying that true scores are not practically relevant.

The third view of true scores comes directly from Lord and Novick (1968). From their perspectives, the concept of a true score is useful both theoretically and practically. Defining a true score conceptually allows the researcher to define errors of measurement. This becomes useful in practice because it allows one to use observed scores to make indirect claims about true scores. Lord and Novick (1968) further distinguished between their proposed way of conceptualizing true scores and the

classical way in which true scores had been conceptualized. Under the classical view, the term “true” score implies a “Platonic” conception of the score which represents the way things “really are” (see Sutcliffe, 1965). This way of thinking about true scores aligns with descriptions from early observational astronomy and early statistics. Although this conception might be useful in some physical sciences where the conditions to be measured can be precisely identified, Lord and Novick (1968) argued that this conception is much less useful in psychology where most theories “are based on unexplicated, inexact constructs” (p. 28).

In presenting an alternative view, Lord and Novick (1968) argued that since the “true” score is an unobservable that is not directly measurable, true scores are hypothetically *indirectly* measurable because they are theoretically related to the person-level propensity distribution of observed scores (which is hypothetically *directly* measureable). That is, the true score is taken to be to the average, or expectation, of infinite and repeated measurements obtained at the person-level. Lord and Novick (1968) argued that it is more sensible, both mathematically and semantically, to define the “true score” as an “expected value.” Thus, they argued for an *operational* conception of true score as an expectation. This point regarding true scores has been widely adopted under test theory today, including under generalizability theory. In modern literature, it is common to use the term “expected value” in place of “true score” so as not to imply a Platonic version of the concept (e.g., Traub, 1994). This being said, Lord and Novick (1968) were clear that they did not mean to imply that the Platonic conception of true score has *no* place in test theory.

> The Platonic concept of true score is not one which should be or is likely to be completely neglected. Indeed, factor analytic theory was originally conceived in Platonic terms. We simply point out here that the operationally related definition of true score [i.e., the “expectation”] which we usually adopt is in some ways a very convenient one theoretically (p. 42).

Lord and Novick’s (1968) views on true scores hold implications for their definition of “error.” They noted that a carefully controlled experiment in the physical sciences has the potential to eliminate most of the error associated with measurements, whereas this task would be much more complicated in psychology. From Lord and Novick’s (1968) view, the error random variable, *E*, can be defined as a “disturbance” due to controllable and uncontrollable factors. The fact that some of the variation

associated with the error random variable can be controlled in the process of setting up a testing environment led Lord and Novick (1968) to argue that "the error random variable and the true score are determined by experiment, not by some hypothetical state of affairs" (p. 38). This argument logically follows from their distinction between Platonic and operational true scores. Thus, in the same way that Lord and Novick (1968) opted to operationally conceptualize true scores as expectations, they also opted to conceptualize the error random variable as the "residual random variable."

3.6. Error in Historical View

The previous chapters have outlined a history of error from its uses in early astronomy and statistics to its adoption under 20th century psychometrics and experimental psychology. Given that this history spanned several centuries, its focus was mainly on the "big picture" uses of the error and true score concepts rather than detailed examinations. From the history described, it is clear that error has played a central role in the development of statistics and its utility in scientific endeavours. In fact, one might argue that the history of statistics is essentially a history of the error and true score concepts, as advances in the uses of these concepts were fundamental to the development of statistics and the dominance of statistical tools in scientific practice. In psychology, the works of early prominent statisticians and probability theorists such as De Moivre, Legendre, Gauss, and Laplace have been central to the advancement of the error and true score concepts within the field. Psychologists have adopted and expanded uses of the random variable model (i.e., true score model under test theory), the method of least squares, and error analysis. Interestingly, psychology has modeled its practices directly after the practices of the physical sciences even though it has dealt with vastly different objects of study. That is, psychology adopted the error concept, originally used to study observable phenomena such as planets, for use with the study of unobservable psychological phenomena such as intelligence and other mental traits. To better understand how this was done, in the next chapter, I examine the biography of the error concept and identify common themes. Based on these themes, I consider several underlying issues related to the uses and meanings of the error concept in psychology.

Chapter 4.

An Analysis of Uses of the "Error" Concept in Psychology and Statistics

It is apparent from the history outlined above that the concept of error took on several different roles and uses throughout the history of early and modern psychology. The goal of the current analysis is to examine these various uses and their implications for the meaning of the error concept. As mentioned previously, "error" as a statistical and methodological concept is technical in its use. Researchers aim to use this concept in a way that is different than its uses in ordinary language. Therefore, I restrict my analyses to uses of "error" within the linguistic domains of statistics and psychological science.

The current analyses are guided by the following 3 overarching research aims:

- **Research Aim 1: Comparison of Error in Psychology and Error in Early Astronomy/Statistics.** In what ways do uses of the concept "error" in early and modern psychology overlap with its uses in early statistics, and in what ways do they differ? Given that psychology is concerned with unobservable human phenomena, did the usage of the concept "error" need to be revised under psychology? If yes, what were the implications of this for the *meaning* of the term? And further, what implications does this have for the study of psychology?
- **Research Aim 2: German vs. British Psychology Traditions.** Did "error" play a different role in the experimental tradition of psychology tied to the early works of Fechner and Wundt than it did in the correlational tradition of psychology tied to the early works of Galton and Pearson? What influences did each of these traditions have on 20th century psychology (including the adoption of Fisher's ANOVA) and what was the impetus of the adoption of the concepts of "error" and "true score" in classical test theory?
- **Research Aim 3: Levels of Analysis.** What are the implications of the *level of analysis* for the conceptual meaning of "error?" Here, I will aim to clarify and distinguish between "error" at the intra-person level and "error" at the inter-person level.

4.1. Analysis Plan

Given that the current project focuses on the study of a technical concept that has been defined specifically for its utility in statistical activities, I conducted my analysis

by focusing primarily on the *applied uses* of "error" in research practice. I investigated how statisticians and psychometricians have described "error" as well as how they have used it in common statistical and psychometric procedures. Thus, my focus in this analysis was on the grammatical usages of the "error" concept, with emphasis being placed on its uses within the statistical/mathematical activities of psychologists. Error is also a concept that is integral to the historical development of mathematical statistics as a discipline. To understand the history of statistics, one needs to have a solid understanding of the error concept. Conversely, to understand the error concept, one needs to have a solid understanding of the history of statistics. Thus, my analysis is structured temporally. I outline historical developments pertaining to the concept of "error" under statistics and quantitative psychology. Finally, I also draw from the qualitative method of thematic analysis (Braun & Clarke, 2006) in organizing my l analysis. Thematic analysis is a general and basic qualitative analytic strategy that involves the identification, organization, and interpretation of themes identified within qualitative data. I adopt strategies from Braun and Clarke's (2006) method of thematic analysis in the following outline of the steps and structure of my analyses.

The initial step in a thematic analysis is for the researcher to familiarize herself with a variety of uses of the topic under study. Given that my current topic of interest is a concept, I aimed to understand different usages of the "error" concept and the contexts within which those usages occur. As such, I familiarized myself with uses of "error" in statistics and psychology by developing a historical biography (i.e., Chapters 2 and 3). The undertaking of such a historical biography of "error" thus constitutes the first familiarization stage of my analysis plan.

Next, I reviewed the history of the error concept to identify interesting features and potential themes. I organized the timeline of each historical period described based on these emergent themes. This was followed by a more in-depth examination of each of the themes to draw connections between themes within the same and different time periods. These connections between themes were then considered in the context of the three general research aims described in the previous section.

Once themes and connections between themes and research aims were established, I made note of interesting points that spoke to each research aim and tied these points to the current theoretical, historical, and philosophical literature. This

allowed me to further build on my themes and to conduct a more detailed analysis situated within relevant literature. These final stages of the analyses were iterative as I continued to further develop the various narratives that emerged from the described history. Ultimately, these analyses led me to several important arguments related to my initial research aims. These are discussed as they relate to each theme in a section entitled, “Integrating Themes and Research Aims.”

In the following results section, I first present a timeline of uses of the concept of error and emergent themes that were identified. I then describe several “superordinate” themes that connect each of the primary emergent themes together. Finally, I draw connections between these themes and my three research aims while drawing from the theoretical, historical, and philosophical literature.

4.2. Results

Table 1 presents an overview of the historical timeline covered in the first section of this project. This timeline acts as a summary of events by highlighting important advances in the history of error chronologically.

Table 1. Central Historical Events and Emergent Themes

Period	Central Events	Emergent Themes
Early Observational Astronomy		
1733	De Moivre develops the exponential function for the bell curve. The variability of this bell curve (normal distribution) is defined as a function of $\sqrt{n}$. This allows for the quantification of uncertainty.	Aim of *accuracy*. True value is the accurate value. Certainty means that the accurate value is captured.
1755	Simpson describes a distribution of errors (i.e., the error random variable). Focus shifts from the average score to the average error. This allows for inverse inference where researchers could reason from observed measurements about true values.	Uncertainty quantified as the standard deviation of the error random variable. Inverse inference from observed measurements to true values.
Late 1700s	Mathematicians continue to consider the properties of the distribution of the error curve. Laplace makes several attempts in defining the error distribution curve, but abandons his work in 1786.	
1805	Legendre develops the method of least squares. He aims to minimize errors to "reveal the truth." The "true" value is taken to be akin to the "accurate" value.	Single observation not useful (observed value is known, error and true score are unknowns). Method of least squares developed to "reveal the truth" based on multiple observations. Goal of eliminating errors.
1809	Gauss uses the normal distribution in the context of least squares. Laplace applies the central limit theorem to the use of the normal distribution. Note that Laplace holds a binary view of true values and errors (true values are permanent, errors are accidental). This is consistent with the prevailing determinist view science of the time.	The normal distribution as "the error curve." Determinist view of true values and errors dominant.
~1830	The method of least squares flourishes in astronomy. It is generally accepted that the error law defines the distribution of errors of calculated averages *and* distributions of errors of observed measurements. Error is attributed to imperfections of measurement instruments.	Distribution of errors of calculated average values *and* distribution of errors of observed measurements defined by error law. Error attributed to instrument imperfections.
Mid-Late 1800s	Astronomers begin to attribute errors to imperfections in measurement instruments *and* human observation. Reaction time is studied as a source of measurement error using the personal equation.	Error attributed to imperfections in human observation (e.g., reaction time). True values exist independent of human observation.

Period	Central Events	Emergent Themes
Early Studies of Human and Psychological Phenomena		
1842	Quetelet studies human traits such as height and birth rates. Quetelet's "*A Treatise on Man…"* describes the average value of an aggregate of values as the ideal. This is done through the introduction of the notion of "the average man."	Quetelet studied social/human phenomena at the inter-individual level using the error law. Society as a social entity. Man as a collective object. Individual differences irrelevant True value as the ideal (average) man Distribution of observed measurements must follow normal curve if no non-accidental causes involved.
Mid- Late 1800s	Fechner establishes the field of psychophysics and uses the concept of error in his studies of sensory phenomena. Fechner holds indeterminist views that contrast with the determinism that had dominated much of science previously. This leads him to an interpretation of "chance" that is independent of random error and not constrained by laws of nature.	Early German psychophysics/psychology concerned with intra-individual variation. Reaction time studied as source of variation rather than as error. Fechner hold indeterminist views. Chance variation independent of random error and unconstrained by laws of nature. True values as real and natural processes. Variations around true values as potentially due to Fechner's interpretation of "chance."
1879	Wundt, highly influenced by Fechner, establishes his laboratory of psychology at the University of Leipzig and relies heavily on factorial design to run studies of *N*=1 (intra-individual variation).	Studies of *N*=1 Reliance on factorial design to "control" conditions.
Mid-Late 1800s	Galton studies inter-individual variation. He is highly influenced by Quetelet's methods but holds the view that inter-individual variation is not merely error, but rather represents important variation in human traits. Galton is not interested in eliminating error; he aims to study it (i.e., he sees error as variation).	Early British psychology (Galton) concerned with inter-individual variation. Galton sees error as meaningful variation. Aim is to study variation, not eliminate it. Studies of variation used to classify people according to traits.
1888	Galton describes the concept of "co-relations" and, in doing so, considers the joint distribution of errors.	Co-relations based on the study of joint distribution of errors
Late 1800s, early 1900s	Pearson mathematically advances correlational analysis. Pearson holds positivist views and does not believe that science should be used to study unobservable causes.	Correlation analysis originates with thinkers who did not believe science should examine unobservables (positivism). Errors of observation are considered fundamental to that which is observed, including observed relations about phenomena.

Period	Central Events	Emergent Themes
Classical Test Theory (CLT)		
1904	Spearman describes the foundations of classical test theory in the context of his work on intelligence. He defines systematic and accidental errors and provides correction formulae for attenuation.	Two kinds of error: random and systematic. Correction for attenuation formula concerned primarily with accidental error. Concerned with minimizing error to obtain the "real" correlation. Error under CLT theorized at the individual level, modeled at the aggregate level. CLT used to study unobservables (e.g., intelligence).
1910	Spearman provides further proof for his correction for attenuation and, in response to criticism from Pearson, he emphasizes the distinction between sampling error and observational error; his focus being on the latter. Spearman also introduces the term "reliability coefficient."	Observational error the focus of CLT. Greater reliability means less error associated with measurements. Reliability theorized at individual level and formalized at aggregate level. Reliability calculated under CLT by splitting tests or duplicating them.
1918/1925	Fisher formalizes the procedure of analysis of variance, emphasizing that statistics operates at the aggregate level. As such, the technique of analysis of variance considers error at the inter-individual level; error at the intra-individual level is not defined.	ANOVA dominates experimental psychology. ANOVA partitions total variability in a variable at the group level. Fisher and aggregate level statistics – error only considered at aggregate level.
1921	Kelley formally defines the concept of a true score for classical test theory. This is defined as the average score of individuals based on repeated and infinite observations.	True scores undefined ontologically for CLT, only technically.
1932	Thurstone summarizes advances in classical test theory, which helps to establish disciplinary norms. The focus of Thurstone's summary is on reliability and validity of measurements. Thurstone emphasizes that chance errors are critical for the field of psychology and that unknown and uncontrollable factors are a "handicap" for psychologists.	Reliability and validity the cornerstones of classical test theory. Psychology, more than any other discipline, is plagued with unknowns.
1945	Guttman clarifies three sources of error: traits, persons, and items. He argues that error should be studied at the person and item levels, but in practice we typically look at error across trials. To reconcile this issue, he provides several lower bounds of reliability.	Observations obtained at aggregate level (across trials) a main problem for classical test theory. Classical test theory moves from the individual level to the group level of analysis.

Period	Central Events	Emergent Themes
1950	Gulliksen summarizes advances in classical test theory. He states that classical test theory is providing confidence that psychology is becoming a science. He emphasizes that the aim of a psychometrician is to achieve accuracy.	The aim of psychometrics is the determination of accuracy of an observed score. Achieving accuracy means establishing a science. Psychological tests have more error associated with them than tests in the physical sciences.
Item Response Theory		
Early 1950s	Psychometricians increasingly begin to critique classical test theory. Scholars such as Guttman and Gulliksen argue that there is no coherent structure to classical test theory and that psychometricians have neglected the issues of sampling error and generalizability.	Classical test theory lacking structure, analysis at item level, and not dealing with problems of sampling error.
1950-1970	Psychometricians (Birnbaum, Lord & Novick, Guttman, Lawley, Lazarsfeld, etc.) develop item response theory, which places focus on items. The notion of "information" is introduced which allows psychometricians to think about varying levels of error at the item level and at differing abilities levels. Precision now conceptualized to vary across the range of possible scores that one might obtain on a test (and across items). Moreover, errors in scores are conceptualized as varying as a function of the value of a trait (i.e., ability).	"Information" under IRT is large when CLT concept of "precision" is large. "Information" can vary by item and ability, "reliability" under CLT is at the test level.
1968	Lord and Novick's book, "*Statistical Theories of Mental Test Scores,"* including Birnbaum's contributions, establishes item response theory. In this book, Lord and Novick also describe three different views of "true scores" in psychometrics: theoretically irrelevant, practically irrelevant, and both theoretically and practically relevant. They argue for the latter view and put forth the notion of an "operational true score" (i.e., an expected value) distinct from the Platonic notion of a true score. They also redefine the error random variable as the residual random variable.	Disagreement in psychometric literature about whether true scores have theoretical or practical value. Distinction made between Platonic conception of true score and operational conception. Operational conception of true score adopted in psychometrics. Operational conception of true score shifts view of "error" to "residual."

Period	Central Events	Emergent Themes
Generalizability Theory		
1972	Cronbach et al. (1972) introduce generalizability theory based on Fisher's analysis of variance. One goal of Cronbach and colleagues is to unite the two methodological traditions of psychology (psychometrics and experimental psychology). Under generalizability theory, random error can be attributed to more than just one undifferentiated error source. This is done using the technique of analysis of variance in which the total variability in an observed score is separated into parts. In this way, generalizability theory also accounts for sampling error.	G theory meant to unite ANOVA and test theory. G theory partitions error sources (differentiated error). G theory addresses sampling error by defining "universes." Under g theory, reliability speaks to the accuracy of generalization. "Sources" of error emphasized differ under ANOVA and test theory traditions.

4.2.1. Connecting Emergent Themes

The emergent themes described in Table 1 are summarized below through several "superordinate" themes that connect each of the primary themes together. Throughout the timeline described above, several strands of thought that connect the emergent themes are apparent: (1) the pursuit and aim of accuracy; (2) the consideration of error over repeated measurements; (3) the consideration of various sources of error; (4) a distinction between "error" and "variation;" and (5) a distinction between "intra" and "inter" levels of variation/error.

4.2.1.1. Aim of Accuracy

In describing the method of least squares, Legendre explained how his technique could be used to reveal or approach "the truth," or the most "accurate" possible result (as cited in Stigler, 1986). Quetelet (1842) described the "average man" as the ideal, and all individual difference was merely viewed as deviation from this ideal. In 1950, Gulliksen described the aim of the psychometrician as that of the determination of the "accuracy" of an observed score. The pursuit of accuracy has been a common and invariable theme across the timeline of the error concept. It is clear from the history of quantitative science that any discipline claiming "scientific" status should uphold a standard of accuracy. However, what is meant by "accuracy" is less clear, as interpretations of the term and methods for deducing putatively accurate results have varied over time and between

disciplines. To understand what it might mean for a researcher to obtain “accurate” observations or “accurate” estimates, one must first understand what it is that the researcher is striving to obtain. In the current context, this means the concept of the “true score” must first be defined in order to determine if differing interpretations of “accuracy” are in large part a result of differing interpretations of the “true score.”

For early observational astronomy, the true value was equivalent to a naturally occurring feature of the solar system, such as the real orbit of a planet as it truly exists. The true value was something “out there,” independent of human observation. However, being that human observation has its imperfections, the true value was not directly observable by humans. To achieve accuracy, then, meant to achieve an observed estimate of the true value that was as close as possible to the actual true value as it truly exists. A completely accurate estimate would be one that is *equivalent* to the true value (i.e., the true value *is* the accurate value). This led to the conception of another important term, “certainty.” That is, how “certain” could a researcher be that they had accurately captured the “true value” in their observations? Certainty was quantified with the introduction of the error random variable. The less variability associated with the error random variable, the more certainty researchers attributed to their observations. This was based on the notion that the closer observations were to the average value (and to one another), the more certain the researcher could be that they were close to the accurate value. Thus, the aim of early observational astronomy was to achieve accuracy in the sense of obtaining the true value, as it truly existed, independent of human observation. True scores were explicitly defined as “accurate” scores.

The concept of “accuracy” takes an interesting turn with Quetelet’s notion of the average man as the true score. Under Quetelet’s (1842) view, the average man is an ideal, and thus, although it may not be represented in reality, it is an ideal to strive for. Quetelet is concerned primarily with the random variable of observations, rather than the error random variable. Under his view, variability of the observed random variable is in a sense “naturally occurring” in that deviations from the average man are a result of nature’s error law. For Quetelet, to strive for accuracy meant to strive to be, for example, the same height or the same weight, as the average man. Unlike early observational astronomers, Quetelet is less concerned with matters of reality as he is concerned with the ideal. Why such a shift in the interpretation of the true value? One reason may be due to the shift in content matter. Whereas astronomers were concerned with non-

directly observable planets, Quetelet was concerned with the measurement of human and social characteristics. The phenomena that were the objects of Quetelet's studies were thus more easily directly measurable (e.g., height, birth rates, death rates) and were also more political in nature. Moreover, Quetelet was concerned with averages over observations of *different* people; whereas, astronomers were concerned with observations of, for example, the *same* planet's orbit over differing points in time.

This distinction between level of analysis and the impact that it has on interpretations of true scores is particularly important to note for the proceeding discussion of accuracy in the study of psychological phenomena. The issue of level of analysis will be examined in greater detail in a following section. However, for the purposes of understanding the true score concept, it is important to note that early investigations of psychological phenomena were conducted at different levels. Fechner and Wundt were interested in making claims that were generalizable to *all* individuals (Lamiell, 2003). This meant that they were not particularly interested in average values obtained from groups, but rather from individuals. If a hypothesis falsified for any single person, then it could not be deemed true for all. Fechner and Wundt's studies thus largely consisted of repeated observations of the same individuals to allow for the positing of general-type claims (i.e., generalizations to populations of individuals). This was, perhaps, also due to the intensive nature of their method of choice, introspection, which consisted of participants observing and reflecting on their own mental states. Such a method required substantial training of research personnel and commitment from participants in order to be implemented. As such, these researchers studied variation at the intra-individual level, and thus, a true score for them would be an *individual* value (Gigerenzer, 1987). For example, the "true" reaction time over repeated measurements obtained from the same person exposed to the same stimulus over time. In this way, true scores under Fechner's psychophysics would have been most similar to their interpretation under early observational astronomy. In contrast, Galton was mainly interested in making claims that were true on average and in studying variations *between* individuals. Thus, he primarily studied inter-individual variation and was focused on a level of analysis that would have been more akin to Quetelet's studies (Porter, 1986). Unlike Quetelet, however, Galton (1888) was more interested in variation around the true score rather than the true score itself. For this reason, discussions of the true score are not an integral part of Galton's work. Perhaps, then, the pursuit of

accuracy for Galton could be interpreted as the pursuit of accurate *variation* rather than the pursuit of an accurate score. This is because Galton did not aim to eliminate variation around an average value; rather, he aimed to understand the variation around it and believed that this variation itself was a naturally occurring phenomenon that was not necessarily bound by the error law (i.e. not due to random error in the way that Quetelet would have interpreted it).

Finally, under 20th century psychometrics we see several important concepts take a central role in the measurement of psychological phenomena. Of note are the concepts of "reliability" and "precision" – both concepts that are tied to the notion of "accuracy" in the classical astronomy sense. However, psychologists are clearly not interested in the same kinds of phenomena as astronomers. Although the subject matter of astronomers is out of reach and not directly observable without measurement instruments; it is debatable whether psychological phenomena are comparable. Indeed, given that Spearman believed that correlational methods could uncover underlying causes of human variation (e.g., true intelligence scores), he also would have likely argued that the phenomenon of intelligence was akin to the orbit of a planet around the sun in that it could be attained indirectly through observed measurements. It is under this way of thinking that psychometricians adopted the notion of a true score and the pursuit of accuracy from the physical sciences. Surprisingly, however, many early psychometricians did not delve too deeply into the meaning of a true score. Although some scholars, such as Spearman (1904a), did indeed discuss ontological and epistemological matters (e.g., Spearman was interested in intelligence ("g") as a real phenomenon and his early works (1904a, 1904b), in which he discussed real values and the pursuit of uniformities in science), others, such as Kelley (1921), incorporated operational definitions of true scores. Moreover, Lord and Novick (1968) addressed the problem of the true score but focused primarily on the issue of whether the concept is theoretically or practically meaningful for psychometricians. They concluded that the true score is both theoretically and practically useful, but that it would be unhelpful to adopt a Platonic view of true scores wherein they are defined as matters of reality as they truly exist. The notion of the "expected value" was proposed to get around this issue of the Platonic true score. By redefining true scores in purely operational terms, psychometricians could bypass the ontological issue of true scores and yet maintain their status as a science in pursuit of accuracy. Importantly, this does not suggest that

psychometricians were necessarily devoid of any ontological commitments, but rather that they focused primarily on describing the operational uses of true values.

4.2.1.2. Error Over Repeated Measurements

The shift in focus from the observed random variable to the error random variable was a pivotal point in the development of methods for achieving accuracy in the field of astronomy. The application of the normal curve to the error random variable distribution not only allowed researchers to combine multiple measurements over time and under varying circumstances, it also allowed them to implement the method of least squares as a tool for minimizing errors (Stigler, 1986). Single observations of phenomena were (and still are) considered inadequate because (1) every single observation has some error associated with it and (2) the amount of this error cannot be deduced from only a single observation, as a single observation contains two unknowns (the true score and error) and only one known (the observation itself). However, with multiple observations, one could use the method of least squares to deduce the level of certainty associated with observations and select a single value that best minimizes the amount of error. Thus, the examination of error over repeated measurements has been a common theme throughout the history of statistical science. As a result, single observations have been deemed insufficient.

The method of least squares was therefore a blessing for the field of astronomy which aimed to combine observations of the same phenomenon obtained under varying circumstances. This method has of course been applied to the study of psychological phenomena, particularly in the experimental stream of research following from Fisher's analysis of variance and the development of regression analysis. Whereas these approaches all consider the error random variable at the aggregate level (i.e., as a result of inter-variation), psychometric theory, especially classical test theory, was primarily concerned with intra-individual variation (i.e., repeated measurements obtained under *identical* circumstances). However, because obtaining repeated measurements under identical circumstances is impossible for psychology, we see an interesting shift in classical test theory wherein procedures are theorized at the intra-individual level, but formalized at the inter-individual level. Thus, in theory, classical test theory is concerned with repeated measurements obtained from the same person in vacuo; however, in practice, it deals with repeated measurements over varying circumstances and from

different people. Unfortunately, as discussed in a later section, this shifting between different levels of analysis has the potential to cause confusion in the interpretation of results and in how we understand the role that error plays in psychological methods.

4.2.1.3. Sources of Error

Where does the "error" associated with observations come from? This has been a third continuing theme throughout the history of the concept. To ask this question is to ask a question that is both ontological and epistemological. That is, one must consider the nature of error sources and how an observer perceives or knows about such sources. The first uses of the "error" concept in early observational astronomy were in a sense quite "practical." Astronomers were interested in obtaining a numeric value that was presumed unobservable or "out of reach." That is, they were interested in physical features of planets, most of which could be represented in terms of extensive quantities such as length or distance. The latter were measurable in the sense that standard units could be applied such that they were "captured" via measurement instruments. Thus, astronomers aimed to observe features of physical objects that were deemed too distant to be measured directly. As a result, they developed instruments (e.g., telescopes) to provide indirect measures of these features; however, these instruments were prone to inaccuracies. That is, no single observation of the desired numeric value could guarantee accuracy. It was expected that there would always be some amount of deviation between the observed value and the desired value. This deviation was thus termed "error." Astronomers adopted a "realist" perspective of their subject matter under which it is assumed that an observer using a telescope to view a planet will have no impact on that planet's existence (Slaney, 2001). The planet is believed to exist "out there," independent of human interference and, so too, the features of the planet that were of interest. The only method of human observation of such a planet would be through human-made instruments, such as the telescope. Given this realist view, "error" appears to have a somewhat "practical" source in astronomy. By this I mean that we can pinpoint several practical reasons as to why an instrument such as a telescope would not provide accurate results. First, the instrument itself may need to be refined and improved for accuracy. Second, human observers are imperfect; they may have slow reaction times in recording observations, or perhaps their eyesight does not allow for accurate observations.

Thus, initially, early statisticians (e.g., Laplace) and astronomers attributed error to random sources associated with imperfections of measurement instruments. Over time, other sources of error related to imperfections in human observation (e.g., reaction time) were also acknowledged (Stigler, 1986). These errors were referred to as "chance" variations in that they occurred randomly (as a product of accidental forces). This could be contrasted with true values that were a product of naturally occurring and permanent forces (under some interpretations, these forces were considered "laws"). As such, early views on sources of error were deterministic. Chance variation due to random error was only viable because of human ignorance and imperfections in measurement instruments. This was the extent to which any sense of indeterminism was accepted. All naturally occurring phenomena in the world was determined (Heidelberger, 1987).

This deterministic view was also adopted under Quetelet's interpretation of the error curve. For Quetelet, even random error could potentially be interpreted as being due to deterministic forces in the sense that variation amongst observations of human phenomena at the group level always followed the same bell-shaped curve (Porter, 1986). Under this view, the normal distribution was a "law." A realist perspective was carried forward in Quetelet's studies of social phenomena. He did not presume that his observations of characteristics such as height, weight, or birth rates would have any impact on his subject matter. He used measurement tools and instruments (e.g., scales, counts) to obtain his observations. Although Quetelet's interpretation of error differed from that of early astronomers in that he believed that errors were deviations from some ideal value, like astronomers, he too adopted a determinist view of his subject matter. Under this view, true values were the products of permanent forces and the only "chance" or probabilistic variation that could occur was random error. However, even random error was also believed to constrained by natural laws.

What was considered to be a source of error under observational astronomy was later considered an interesting phenomenon in need of study. Here, I am referring to the phenomenon of reaction time, which would become a central focus of Fechner's studies in psychophysics (Stigler, 1986). For Fechner, reaction time was a source of *variation* rather than a source of *error.* In fact, Fechner also held views of chance variation that went against the dominant determinist way of thinking at the time (Heidelberger, 1987). Under the determinist view, chance variation was akin to what was considered as "random error." However, for Fechner, chance variation was independent of random

error and unconstrained by laws of nature (Heidelberger, 1987). Thus, under Fechner's view, we see a shift from a deterministic view of variation to one that allows more room for indeterminism. Moreover, although Galton did not promote indeterminism in the same way as Fechner, he differed from Quetelet in his examination of error and thereby also in his interpretation of its source. Whereas Quetelet was interested in aggregate-level error insofar as it allowed him to compute estimates of the average man, Galton was interested in error as a means of studying variation. Thus, differences in traits between people provided a meaningful source of variation for Galton.

Finally, prior to 20th century psychometrics, scientists often discussed "accidental" vs. "non-accidental" errors. However, this distinction between two kinds of observational errors was highlighted under classical test theory, particularly with Spearman's discussion of "systematic" vs. "random" errors. The latter is what was most concerning for Spearman. Systematic errors can largely be controlled through methodological design; however, there will always be some amount of random error associated with an observation. This distinction between systematic and random error was wedded to some extent on how well the complexity of sources of error was worked out under early test theory. Indeed, in the early literatures of classical test theory, one does not find many discussions of philosophy, or the nature of test theory concepts. This is not to suggest that this literature is completely lacking, as many early psychologists did write about such matters. However, early test theorists seemed to be much more concerned with developing methods and uses of procedures than contesting the philosophical groundings of such procedures. This is also true of later developments under test theory such as item response theory and generalizability theory. Although the error concept is expanded under these frameworks, there is little discussion of its ontological or epistemological implications. Under each of these frameworks, the use of the error concept is mostly operational.

4.2.1.4. "Error" vs. "Variation"

Regardless of the source from which an error value is theorized to derive, for many of the early thinkers described above, an important aim of science was to eliminate observational errors. However, this aim only holds if deviations around a central value are indeed interpreted as error (i.e., inaccuracies or deviations). Thus, it is important here to consider the difference between the concepts of "error" and "variation."

The idea that deviations from a central value might in themselves be interesting topics of study does not appear to have been given much attention until the mid-19th century, particularly with the works of Fechner (i.e., studies of reaction time) and Galton (i.e., studies of co-relations). Indeed, astronomers were interested in eliminating errors through the method of least squares, and Quetelet was similarly interested in eliminating errors to reveal the average man. However, with Fechner and, especially, Galton, we see a shift in the interpretation of error in which deviations from a central value (e.g., the mean) are potentially viewed as meaningful, and oftentimes, naturally occurring, variations. Although Fechner did consider some error sources (e.g., reaction time) to be meaningful topics of study, he did not aim to study variation in the way that Galton did. Moreover, it is important to keep in mind that Fechner was dealing with variation at the individual level; whereas, Galton was examining variation in scores across different people. Thus, Galton used variation as a method for examining individual differences and correlation analysis was in part developed as a way of examining this variation. Although Spearman and the early classical test theorists adopted correlational methods from Galton and Pearson, they were more interested in the elimination of errors than in the study of variation. Thus, the interpretation of deviations as error vs. variation appears to rely largely on one's scholarly purpose. If one's aim is to describe changes in individual human characteristics, perhaps over time, or to study individual differences in a trait, then deviations from a central value will likely be sought as sources of meaningful change. However, if one's aim is to "uncover" a specific (and unobservable) value, then elimination of "noise" surrounding that value (i.e., error) will be a main goal.

4.2.1.5. "Intra-" vs. "Inter-" Levels of Variation/Error

All the above-mentioned themes are in some way related to the issue of level of analysis. Indeed, a common issue throughout the history described above is the problem of whether researchers are dealing with repeated observations of the same or different object(s)/event(s) under identical or varying circumstances. The notions of intra- vs. inter- individual levels of variation are specifically tied to the study of persons. The "intra" level describes a scenario in which multiple observations are obtained from the same person with respect to the same phenomenon, while the "inter" level describes a scenario in which multiple observations are obtained from different people with respect to the same phenomenon. The "intra" level also has two different "sub" levels, the first involving the scenario in which observations are obtained in vacuo (i.e. under non-

varying/identical circumstances), and the second involving the scenario in which observations are not obtained in vacuo (i.e., under varying circumstances). Thus, in total, there are 3 different potential levels of analysis that can be considered when studying characteristics of people. Moreover, a different central value, or mean, will be of interest at each varying level. At the inter-individual level, the central value of interest is the average score on a trait across different people. At the intra-individual level with observations in vacuo, the central value of interest is a person's "true" score, which is the fixed expected value (mean) of the propensity distribution for that individual and a given measurement operation. At the intra-individual level with observations *not* obtained in vacuo the central value of interest is a person's average score from a given measurement operation across time or varying situations. In contrast to true scores, average scores can potentially vary, depending on the variation of circumstances under which a phenomenon is observed.

These characterizations of levels of analysis are based on observations taken from people. For early astronomers, the objects of study were often planets. As with people, it would be impossible to obtain intra-level observations of planets in vacuo; however, intra-level repeated measurements under varying circumstances were possible and were typically what astronomers were interested in obtaining. The circumstances that would often "vary" for these researchers included factors such as time and changes in the human observer. For astronomers, these factors were easier to consider as "noise" because the location of a planet does not substantially change within minutes and it was presumed that an observer typically would not have a direct impact on the location of a planet (moreover, the notion of a "location" in space is not necessarily objective; rather, locations are defined by astronomers relative to some reference point, e.g., another planet, the sun). In some ways, the subject matter of astronomy made it easier for astronomers to discount extraneous factors that might influence their observations.

Unlike astronomers, Quetelet dealt with inter-level observations and was not very concerned with distributions of intra-level observations. The central value of these inter-level observations obtained under varying circumstances was interpreted by Quetlet (1842) as a true value, and deviations from this central value were interpreted as error. Thus, there was, for example, only one ideal height for man and the average calculated over individual heights was a representation of that ideal value. Like Quetelet, Galton

also dealt with inter-individual observations; however, he interpreted the average as just that – the average value, i.e., in a strictly arithmetic sense. Galton was more concerned with variation around the average than the average value itself (Porter, 1986). Moreover, in his correlational analyses, he was interested in how observations on two different variables of interest varied together. Unlike Quetelet and Galton, Fechner and Wundt examined distributions of intra-individual variation. Their studies were, in essence, tests of "repeated measures" and they attempted to "control" for extraneous factors that might affect a participant's performance through factorial design. However, their observations were by no means obtained "in vacuo," nor were their inferences constrained to the individual cases they studied, as was often the case in astronomy.

The issue of level of analysis became increasingly complicated in 20th century psychology. Fisher (1925) was quite clear that ANOVA was meant to be used and interpreted at the aggregate (inter-individual) level and that average values obtained through ANOVA did not speak to individual cases, nor were they to be interpreted as "true" values in the same sense as in classical astronomy. In contrast, the core of classical test theory is the person-level distribution of observations that is theoretically meant to be obtained in vacuo. It is the intra-individual variation between measurements taken form the *same* person under the *same* circumstances that is foundational to classical test theory. However, test theory is carried out at the inter-level of observation partly because intra-individual level observations are impossible to obtain in vacuo (i.e., there is always *at least* the factor of time to take into consideration). This is a central problem of classical test theory and the main impetus for the development of different kinds of "reliability" which are, in essence, different methods for getting around the issue that single observations are obtained from different people rather than multiple observations from the same person.

4.2.2. Integrating Themes and Research Aims

As outlined in the above sub-sections, my historical review of the concept "error," revealed several emergent and re-current overarching or "superordinate" themes. In this section, I consider these themes in the context of my initial research aims. In doing so, I draw from scholarly works in the history and philosophy of science to ground my results within the existing literature. My examination is rooted in several "arguments" regarding uses of the error concept that are based on my analyses.

4.2.2.1. Research Aim 1: Comparison of Error in Psychology and Error in Early Astronomy/Statistics

My first aim is to draw comparisons between how the concept of error was used in early astronomy/statistics and in traditions of psychology. To this end, I emphasize similarities and differences between the objects of study within these disciplines. It is apparent that usages of the error concept were revised with scientific psychology, particularly because of the special nature of its subject matters. In turn, this has had implications for the *meaning* of error as well as how the study of psychological phenomena is approached today. Based on my analyses, I argue that emphasis was placed on the operational uses of the concepts "error" and "true score" in 20th century psychology, while ontological and epistemological questions related to the uses of these concepts were often bypassed. This meant that the complexities associated with the ontological natures of psychological phenomena were often reduced to issues of statistical error rather than directly addressed.[4]

There are two important changes that coincide with the adoption of the error concept under psychophysics, and ultimately, psychology: the nature of the objects of study, and the way in which objects are "measured." Within astronomy, the objects under study were commonly of the physical form, or, at the very least, objects that could be "pointed to." That is, if an astronomer is examining characteristics of a star, she can presumably "point to" that star on each repeated observation. This was difficult to do with phenomena such as reaction time, intelligence, or personality. Moreover, whereas astronomers use instruments such as telescopes to obtain observations and/or measurements of phenomena, the measurement of mental phenomena using so-called "instruments" is much more complex. For this reason, test theory treated its "tests" as the instruments of measurement; although, it is arguable whether a "test" can be considered a measurement device (Michell, 2004; Swijtink, 1987). Given the complexity of measurement in psychology, the psychometric literature of the 20th century was largely focused on developing a theory of mental testing rather than determining the philosophical implications of adopting statistical procedures. Indeed, Lord and Novick (1968) stated that the development of a theory of mental testing is of upmost importance

[4] The complex nature of psychological phenomena is a large and thorny topic that goes beyond the focus of the current project. As such, I have not thoroughly examined it here. However, for the current purposes I wish to emphasize only that such complexities were often not discussed in the psychometric literatures of the 20th century.

for psychology because it is an area of study that deals with unobservable phenomena.[5] They argued that because psychological objects of study are unobservable, psychological researchers will deal with larger amounts of error associated with their measurements than that which is found in the physical sciences. The random variable model was adopted from the physical sciences and directly applied to the study of psychological phenomena, such that "differences in scores [on multiple administrations of a test] [were taken to] represent the failure of the measuring instrument to do what we wish it to do" (Lord & Novick, 1968, p. 13). Psychological tests were perceived of as being the "instruments" of psychology in the same way that a telescope was an instrument of astronomy. Differences in scores obtained using tests were interpreted as being due to random error in the same way that random error in early astronomy was attributed to imperfections in measuring instruments. The model of measurement in astronomy is thus applied to the study of psychological phenomena with the caveat that psychological measurements will carry greater error than measurements obtained in the physical sciences due to their "unobservable" natures (Gulliksen, 1950).

There is no doubt that the perception that there are large amounts of error associated with the measurements of psychological phenomena was a primary concern of psychometricians throughout the 20th century. A great deal of focus was placed on handling different types and sources of error, as is evidenced by the developments of item response theory and generalizability theory (Markus & Borsboom, 2013). In doing so, many psychometricians chose to emphasize the operational uses of the concepts "error" and "true score" in their writings (e.g., Kelley,1921; Lord & Novick, 1968). This was perhaps a tactical move that allowed psychometricians to bypass the philosophical complexities that came along with studying human phenomena understood as inherently "unobservable" and thus only indirectly measureable. However, this does not imply that psychometricians held no ontological commitments regarding their objects of study. On the contrary, it is quite clear that many psychometricians treated their objects of study as existing independent of human influence, such that mental tests could be seen as being administered objectively to participants as a means of uncovering underlying causes of

[5] Although I make reference here to the term "unobservabilty," I acknowledge that this term may not be the most accurate way of representing the private nature of consciousness and associated concepts. However, given that the current work is concerned with how error has been conceptualized within psychology and because the term "unobservable" is often how psychological phenomena have been framed within the historical literature, it is important to capture that fundamental assumption here.

behaviour. That is, many psychometricians in the 20th century were ontologically realist. For example, this a perspective that was outlined in the early works for Spearman (1904a, 1904b) and that was illustrated in Birnbaum's (1968) description of latent traits:

> In any theory of latent traits, one supposes that human behaviour can be accounted for, to a substantial degree, by isolating certain consistent and stable human characteristics, or traits, and by using a person's values on those traits to predict or explain his performance in relevant situations (p. 537).

Psychometricians have generally held a realist perspective on the nature of their objects of study, although they tended to emphasize operational definitions of the "error" and "true score" concepts in their works. For example, psychometricians such as Lord and Novick (1968) and Cronbach et al. (1963) explicitly distanced themselves from a "Platonic" notion of true scores and instead proposed operational definitions of true scores as "expected values." Again, as alluded to above, this was perhaps one way of bypassing the complexities that came along with studying psychological phenomena while still applying the classical model of error theory from the physical sciences to psychology. However, the question of whether a model built around the physical sciences can be directly applied to the social science remains open. This is, in fact, a question that has been asked by philosophers of science for centuries. There are several views, following from the works of scholars such as Dilthey (2002), Gadamer (1960), Heidegger (1927), and Kuhn (1998) that share the idea that the objects of study in psychology carry cultural, social, and historical meanings. From this perspective, humans are self-interpreting creatures and understanding human behaviour requires interpretation. There are several interpretive layers in the study of psychological phenomena that have not been directly addressed in the way that psychometricians have adopted the error model of the physical sciences. In a following chapter, I examine uses of "error" under qualitative research methods, a tradition of research that encourages explicit acknowledgment of these interpretive layers.

A final comparison among uses of error in early astronomy and statistics and modern psychology is related to the issue of determinism. Although psychology adopted probabilistic methods that were popularized towards the end of the 19th century by Fechner and Galton, it appears to have adopted a classical deterministic way of viewing phenomena under study. Laplace, a strong determinist, held the view that all things in life are determined by laws of nature and that all occurrences have an underlying cause

(Kruger, 1987). Similarly, in 1904, Spearman remarked, “all knowledge – beyond that of bare isolated occurrence – deals with uniformities” (p. 72) and that correlational analysis could be used to uncover “hidden underlying cause[s] of variations” (p. 74). Moreover, Lord and Novick (1968) were explicitly clear in their discussion of mental test theory that although the true score model is probabilistic, its interpretation under psychometrics is entirely deterministic. All error in a model was attributed to human ignorance and/or the design of the experiment. Gigerenzer (1987) has argued that 20th century psychology, particularly that of the 1920s-1950s, used probabilistic thinking to achieve the ideals of determinism and objectivity. Here, objectivity is meant in two ways. First, probabilistic models allowed psychologists to create the illusion that psychological objects of study were independent of the observer. As Daston and Galison (2007) aptly stated, “to be objective is to aspire to knowledge that bears no trace of the knower” (p. 17). Gigerenzer (1987) argued that statistical inference, particularly as embedded within the method of ANOVA, played an important role in constructing this illusion for psychologists. Second, probabilistic thinking also allowed psychologists studying mean differences using techniques such as ANOVA to create an illusion of objectivity in the sense that individual differences at the group level were interpreted as merely “error” under this method.[6]

Thus, psychology adopted the random variable model and the concepts of “error” and “true score” from early studies in the physical sciences. However, it faced the challenge of applying measurement to unobservable psychological phenomena. This issue was often reduced to the dealing with large amounts of error, which meant that a central task of psychometrics was the development of a test theory. Although many psychometricians likely held strong ontological positions regarding the natures of psychological objects, many chose to stick to operational definitions in their descriptions of true scores and errors. This was likely a tactical move (either implicitly or explicitly) to bypass complexities associated with the philosophical implications of applying the error model to the study of unobservable phenomena. Finally, although psychology adopted probabilistic methods as the central tools of its discipline, it held onto the deterministic views of early astronomy and statistics.

---

[6] Many psychologists in the 20th century were still interested in the Galtonian tradition of studying individual differences and not all psychologists treated such differences as merely error (see Revelle, Wilt, & Condon, 2011). However, statistical techniques such as ANOVA were commonly applied to the study of psychological phenomena in a manner such that individual differences were treated as error.

4.2.2.2. Research Aim 2: German vs. British Psychology Influences on 20th Century American Psychology

> Psychology is proud of its laboratories, with their apparatus for careful experimentation and measurement. It is proud also of its array of tests for measuring the individual's performance in many directions. It is pleased when its data can be handled by mathematical and statistical methods (Woodworth, 1929, p. 7-8).

The second aim of the current analyses is to understand the role that the error concept played in the research practices of Fechner and Wundt in Germany and Galton and Pearson in England. More specifically, I intend to explicate the influences that each of these traditions had on uses of error in 20th century psychology. In 1957, Cronbach pointed out that psychology had been divided into two disciplines: experimental and correlational psychology. The former "studies only variance among treatments" while the latter "studies only variance among organisms" (p. 681). The tools of psychometricians, particularly factor analysis (and other latent variable models), would fall into the tradition of correlational psychology, while the use of ANOVA to examine differences between treatment and control group means would fall under the tradition of experimental psychology. The early works of German and British scholars would have a profound influence on 20th century American psychology, and, depending on the methods used by the researcher (i.e., whether experimental, correlational, etc.), this influence would take on different forms. Here, I begin by describing the influences of German and British psychologies on 20th century experimental psychology and then describe the impact of correlational methods specifically on the area of psychometrics.

American experimental psychology borrowed aspects of the probabilistic methods used in both German and British psychology traditions. Danziger (1987) referred to the combination of these two traditions as a "Neo-Galtonian" approach under which the factorial design model is combined with correlational methods for treating variability. The aims of Fechner and Wundt in the uses of error models were much like the aims of early astronomers: to eliminate error and attain accurate estimates of true scores at the individual level (although they also aimed to potentially generalize results to populations of individuals). To this end, the error law was applied as a "Calculus of Error" (Danziger, 1987, p. 39). This can be contrasted with the goals of Galton and Pearson who both aimed to examine variation in true scores (i.e., individual differences), and therefore, to examine error at the inter-individual level. Thus, the use of error models

in this latter tradition can be referred to as a "Calculus of Exploration" (Danziger, 1987, p. 39). Twentieth century experimental psychologists adopted German design and experimentation strategies which were then used in conjunction with aggregate-level statistical tools (i.e., ANOVA, correlation, regression). The amalgamation of these two traditions can be seen, for example, in experimental psychology's uses of "treatment groups." Galton studied naturally occurring phenomena and did not apply manipulations or experimental controls to his subjects. A disadvantage of this method is that it does not allow for causal claims. However, by implementing the notion of a "treatment" group, 20th century experimental psychologists combined the structure and control that would be found in a Wundtian laboratory with the aggregate-level analytical techniques of Galton and Pearson (Danziger, 1987). Thus, experimental psychologists applied error models as both a calculus of exploration (to observe variation at the group level) and a calculus of error (to make causal inferences).

For psychologists in the 20th century, the attainment of true values through experimentation and variation analysis was indeed an important goal; however, psychometricians in the 20th century also held another, and perhaps more important, goal. They aimed to examine error to give credibility to their measurement instruments which consisted primarily of tests or questionnaires. As such, a related aim of psychometricians was to provide evidence that psychology was a "legitimate" science. Indeed, a common way of thinking in the early 20th century was that science *required* measurement (Michell, 2003). Without true measurement, psychology could not claim the status of a true science, and thus, the primary goal of psychometricians was to develop a theory of mental testing in order to allow for the "validation" of psychological tests to show that psychological phenomena could be measured. Unlike early astronomy and psychophysics which focused on eliminating errors and refining measurements to increase precision, 20th century psychology had the additional task of showing that attributes of psychological phenomena were amenable to measurement.

Indeed, for psychologists in the 20th century, establishing a system of measurement was a *necessity* for the progression of the discipline. As Lord and Novick (1968) stated, psychological concepts were believed to be "*defined through* measurement procedures" (p. 16). However, given that psychology dealt primarily with "unobservable" phenomena, it had to "prove" that such phenomena could indeed be measured. To this end, correlational analyses were key. Spearman's uses of correlation

to uncover hidden and underlying causes of variations provided psychometricians with a method for arguing for the possibility of indirect measurement of unobservable phenomena. In particular, Spearman developed factor analysis, a method that is contingent on the notion of conditioning on a factor (i.e., as conceived of as a latent trait) such that observed correlations among scores on two or more measures disappear, which Spearman interpreted as evidence of the common causal source of g (i.e., the trait of general intelligence) for all psychical abilities. Thus, correlational analysis, which was initially used by Galton and Pearson to study natural variation, was ultimately adopted by psychometricians as a method for uncovering hidden causes.

As previously discussed, the use of correlational analysis to uncover hidden underlying causes was in direct conflict with the philosophical beliefs of Pearson, who did not believe that science could be used to study unobservables (Gigerenzer, 1987). However, by the mid-20th century, psychologists would come to adopt a representational view of measurement wherein it is presumed that a homomorphic relationship exists between empirical and numerical relational systems (Berka, 1983). As such, representationlist views commonly involve a form of numerical mapping that is taken to represent quantities of psychological attributes (Markus & Borsboom, 2013). In psychology, Stevens' (1946) proposed a variation of a representationalist view of measurement by asserting that measurement amounted to the assignment of numbers to objects according to some pre-defined rule. This definition of measurement has been the accepted definition in psychology since it was first described by Stevens (1946). The application of measurement (or numerical representation) to unobservable psychological phenomena presupposes the ontological status of such phenomena. Specially, it presupposes that such phenomena exist in, to some degree, in the same manner as the phenomena of the physical sciences (Krantz, 1991). Michell (2003) argued that psychologists have adopted a naïve realist ontological perspective of the attributes of psychological objects. He suggested that by adopting a definition of measurement that is purely operational in nature (i.e., Stevens' previously described definition), psychologists have been able to bypass the question of whether or not psychological attributes actually are quantitative. That is, psychometricians have rarely questioned whether psychological attributes *can* be measured; rather, they have focused on developing a theory for determining *how* they can be measured. Hence, correlational analysis has been adopted as a means of validating measurement instruments (i.e., tests). Thus, like Galton, 20th

century psychometricians used correlational analysis to examine variation; however, unlike Galton, they interpreted correlations between observed variables as being a function of hidden causal factors.

In sum, 20th century psychology took influence from the structures, methods, ideas, and aims of both early German and British studies of human phenomena. How it incorporated various influences largely depended on the goals of researchers (i.e., experimental vs. correlational). In 1957, Cronbach not only argued that psychology was divided between experimental and correlational traditions but that psychologists should seek to bridge this divide such that researchers could simultaneously examine variation among treatments and organisms. In fact, when he and his colleagues proposed generalizability theory based on ANOVA, he was, in part, attempting to bridge the two traditions of psychology. This was done mainly by articulating a method of test analysis that allowed for multiple sources of variation to be considered at the same time (i.e., the partitioning of variability associated with an observation into "parts"). This approach focuses on explicating the *source* of variation (i.e., due to conditions/treatment effects or due to individual differences); however, one must also consider the *level* of analysis or the level at which repeated observations are obtained (i.e., at the person or group level) when considering how 20th century psychology adopted the error model. The following section deals with this final issue.

4.2.2.3. Research Aim 3: Levels of Analysis

My third aim is to examine the implications of the level of analysis for the conceptual meaning of error. My analyses showed that specification of the level of analysis when considering error distributions is important for understanding interpretations of the concept. Although psychometrics, particularly classical test theory, is theorized first at the individual level, practical applications of the theory are conducted at the aggregate level. Moreover, Fisher's ANOVA technique is both theorized and computed at the aggregate level. Thus, the primary statistical tools of 20th century psychology operate at the aggregate level. This is interesting to note because psychology is a discipline concerned in large part with individuals. Based on my historical examination, I argue that part of the reason for this confusion is the transition from individual-level distributions to group-level distributions in psychometrics, as well as the dominance of the aggregate-level approach of ANOVA in experimental psychology.

In both cases, individual-level analysis is abandoned for group-level analysis. I further argue that although these aggregate-level tools are most commonly used in psychology, interpretations of aggregate-level observed score distributions are made as if they speak to general-type claims. As such, psychologists have adopted Fisherian and Pearsonian methods of statistical analysis, and yet use a variation of Queteletian theory in interpreting results based on such analyses. Given this, it is interesting that Quetelet is not typically acknowledged as an important figure in the history of psychology (Jahoda, 2015). This may be an indicator of the level of confusion that exists regarding level of analysis within the field.

As described previously, an important theme in the history of error is the utility of repeated observations over the use of a single observation. The method of least squares provided astronomy with a way of using multiple observations to deduce true score estimates. This also allowed researchers to consider two random variables: the observed random variable and the error random variable. The confusion that appears to have plagued psychology in the 20th century (and certainly still today) is not between these two variables, but at the level at which they are produced. That is, in psychology, one can consider both types of variables at the person (intra-individual) level *and* at the group (inter-individual) level. As Lord and Novick (1968) explained,

> The psychologist often wants to test a whole group of individuals at one time and to make inference about them individually and in relation to the group. The logical and statistical problems in making inferences simultaneously about all individuals in a group introduce many complexities (p. 14).

They added that making inferences about individual events is often easier in the physical sciences because a measurement can be repeated more than once or twice (Lord & Novick, 1968). In contrast, psychological events cannot be measured more than once or twice due to numerous factors including practice effects, changes in psychological phenomena due to time lapse, and participant fatigue. This issue, regarding the lack of repeatability of measurements, plagued psychometricians and experimental psychologists throughout the 20th century (and continues to do so today). For example, classical test theory, including the notion of "reliability," is theorized at the individual level (i.e., intra-individual distributions of hypothetical in vacuo measurements on a single variable); yet, in practice it can only be estimated at the group level. This has led to tests designed to measure individual-level ability being developed and assessed based on

aggregate-level statistics. Similarly, experimental psychologists often want to make inferences about individual behaviour based on aggregate-level analyses such as ANOVA. This confusion in the field of psychology, between "aggregate-type" and "general-type" propositions, was explicitly defined in 1967 by Bakan. He noted that much of psychology is concerned with making general-type propositions that are true of every individual. However, the conventional statistical tools of psychology operate at the aggregate-level; therefore, inferences made based on the uses of these tools can only speak to that which is true of a group of individuals when considered together as a class (i.e., "on average").

Lamiell (2003) has explored the issue of aggregate vs. individual levels of analysis through a historical perspective, noting that researchers such as Fechner and Wundt were interested in person-level distributions (i.e., intra-individual variation) because they aimed to make general-type propositions that were true for all. The works of Galton and Pearson, however, were conducted using group-level distributions of observed variables (inter-individual variation) because these scholars were interested in studying average differences across groups. As previously discussed, the aims of early 20^{th} century psychology were similar to those of Fechner and Wundt in that the focus was on individuals, yet the statistical techniques adopted to address those aims were mainly borrowed from Galton, Pearson, and Fisher. Gigerenzer (1987) argued that one reason for the rise of this "Neo-Galtonian" way of conducting science concerned the level at which the random variable model based on true scores and errors could feasibly be applied to psychological phenomena. In particular, the random variable model dictates that repeated measurements *must* be independent of one another. As previously noted, this is nearly impossible to accomplish in studies of psychological phenomena. Thus, Gigerenzer (1987) argued that psychology had to abandon the individual in favour of inter-individual differences to be able to adopt the random variable model in a way that would be flexible and practical.

Finally, psychology's reliance on aggregate-level error analysis to address individual-level research problems is reminiscent of a "Queteletian" application of the error law. Recall that Quetelet adopted the use of the error law from astronomy, which was concerned with observations of the same event, and applied it to the study of human and social phenomena observed from *different* people. He further applied the interpretation of the error law at the intra-level to the inter-level. That is, Quetelet was not

concerned with individual differences, and in fact only viewed such differences as error. His goal was to minimize error to obtain the true value – the ideal average man. This true and ideal value was meant to represent how matters in nature *ought* to be (Porter, 1986). In this sense, Quetelet was using aggregate-level statistics to make general-type propositions about all individuals. However, the difference between Quetelet and contemporary psychology is that Quetelet promoted a philosophy that supported his unique interpretation of group-level observed random variables. For Quetelet, individual differences did not matter because they were merely error. It would be difficult to find a psychologist today (or in the 20th century) who adheres to such a view. Nonetheless, psychologists have adopted a form of "Queteletian" interpretation of inter-individual error in that they abandoned intra-individual error in pursuit of group-level true values and then used the results of such analyses to make general-type inferences.

4.2.3. Conclusions

Since its beginnings, the field of psychology has modeled its practices after the physical sciences. Like any good science, it values accuracy and the use of precise and rigorous methodology. Psychology has benefited from the random variable model, particularly the conception of errors over repeated measurements, both at individual and group levels. Psychologists have been interested in studying intra-individual variation and inter-individual variation, although, as we have seen, the uses and interpretations of these two levels of analysis have often been conflated. Nonetheless, "error," whether treated as "noise" or "variation," has been a central concept in psychology's pursuit of accurate representations of psychological phenomena. The earliest scholars to study psychological and human phenomena – Fechner, Wundt, Galton, Pearson, and even Quetelet – each discussed the philosophical (ontological, epistemological) bases of their methods choices. The concept of error was explicitly conceptualized in relation to the aims of these researchers. With the advent of 20th century American psychology, an amalgamation of earlier practices was adopted; however, emphasis was placed on practicality and measurement. The concepts of error and true score were often described in purely operational ways. For example, Lord and Novick (1968) re-defined the former as the "expected value" and the latter as the "residual." Moreover, aggregate-level analyses were adopted to answer questions about individuals.

What implications have these practices and uses of the error concept had for the field of psychology in modern times? The techniques that early American psychometricians and experimental psychologists adopted are still widely used today. Factor analysis, item response theory, and generalizability theory are the dominant methods of test development, while ANOVA and regression analysis are the favoured statistical techniques of experimental psychologists. In turn, psychology is still concerned with achieving practical results and describing the complex relationships between groups and individuals. In many ways, the general aims of psychology have remained largely the same, as have its methods. Moreover, psychology has become even more distant from philosophy, a separation that was feared by many early psychological scholars, such as Wundt (1913). This was perhaps fueled by the emphasis that many psychologists, such as Stevens (1946), placed on operational definitions and practicality, as well as on psychology's desire to mimic the physical sciences (see Woodworth, 1929). While taking a practical approach to research may be helpful in producing results, it can also lead to a form of philosophical agnosticism wherein the meanings of key statistical concepts are taken for granted. For example, one can adopt an operational definition of "true scores" as "expected values" (i.e., the average of repeated and infinite observations); however, questions regarding the nature and attainment of true scores, i.e., what they represent (if anything), if they are devoid of human influence, if they can be obtained through observation, remain. Depending on how one answers these questions, the interpretation and uses of true scores in research practice will vary. Moreover, the interpretation and use of "errors" will also vary. Considering that psychologists are interested in mainly subjective phenomena that carries several interpretive layers, perhaps it would benefit psychologists to consider such philosophical questions prior to embarking on research projects.

What might psychology look like if philosophical assumptions were more explicitly acknowledged and questioned in the methodological practices of researchers? One can only speculate that it would open the door for exploration of diverse methods and a greater acceptance of pluralistic approaches to research. Recently, the domain of qualitative research methods has been gaining interest and use within the field. Although qualitative methods have dominated other social sciences, they have been largely dismissed in mainstream psychology. An interesting aspect of the qualitative research domain is the promotion of consideration of the theoretical and philosophical founding of

methods. Given this emphasis, what role might the notion of “error” play in this research domain, if any? This question is explored in the next chapter.

Chapter 5.

Error, Reliability, and Qualitative Methodology

Thus far, I have described how the concept of error played a central role in the development and uses of statistical tools in psychology. I have also discussed philosophical assumptions accompanying uses of the error concept as well as the philosophical implications of its use. Although not always made explicit, the random variable model is commonly used in conjunction with an underlying belief in objectivity and a realist ontology. This includes the notion that objects of study have an objective existence independent of observer influence. Moreover, as described in the previous chapter, many psychometricians and experimental psychologists in the early 20th century emphasized the operational uses of quantitative concepts rather than the philosophical implications of methodological tools. Today, quantitative research methods founded on the basic concepts of error and true scores continue to be dominant in the field of psychology. However, over the past several decades, a movement has arisen in which many psychological researchers are exploring the use of qualitative methods, i.e., methods not based centrally on the examination of numerical information using statistical tools. What's more, the qualitative research tradition *promotes* reflection on theory and the ontological and epistemological implications of methods choices. This is in contrast to the tradition of quantitative methods in psychology in which philosophical assumptions are much less commonly acknowledged.

5.1. Qualitative Research in Psychology

It is important to note that although various forms of qualitative methods are referred to collectively under the title of the "qualitative research tradition," there is a great deal of diversity within and among uses of different qualitative research techniques and the theoretical and philosophical beliefs underlying them. Nonetheless, a trend of growing interest in the general "area" of qualitative research can be mapped in the field of psychology.

Although qualitative methods have not traditionally been highlighted in the research practices of psychologists, such methods have, in fact, been used since the

very beginnings of modern psychology. For instance, Wundt was an advocate of qualitative methods and believed that there should be close ties between psychology and philosophy (Wertz et al., 2011; Wundt, 1913). Wundt used qualitative methods in conjunction with quantitative experiments and reported several qualitative studies in "Volkerpsychologie," a 10-volume piece that emphasized cultural, social, and historical perspectives on the study of people (Brinkmann, Jacobsen, & Kristiansen, 2014; Wertz et al., 2011). Many prominent figures in the early history of psychology used qualitative methods, although these methods were rarely explicated or explored as analytical techniques. For example, Sigmund Freud (1965) used interviews and first-person accounts to capture the meanings of dreams. Qualitative methods were also used by William James (1902) to examine religious and spiritual experiences and Lawrence Kohlberg (1981) to investigate moral reasoning. Moreover, Jean Piaget (1932) is well known for his qualitative studies of child behaviour. These authors did not refer to their works as falling in the domain of "qualitative research methods," nor were such methods ever formally explicated under the framework of a qualitative methodology. Nonetheless, uses of non-numerical based research techniques were common with many prominent psychologists in the early 20th century. In 1942, Gordon Allport put forth an argument that psychology *needed* first-person qualitative data. However, he expressed concern that psychologists often employed qualitative methods in an uncritical manner. Allport (1942) encouraged psychologists to explore the uses of qualitative methods and to approach their objects of study from multiple perspectives and methodological orientations. In his view, multiplicity in method-use was an important aspect of ensuring the validity of knowledge claims.

Despite the promotion of qualitative methods by key psychological figures, uses of such methods have often been marginalized in the field of psychology. Several scholars have argued that this has been due to psychology's struggle for acceptance as a science in which analytical tools based on mathematics and statistics are deemed more rigorous and objective than qualitative research tools. For example, Lamiell (2013) described psychology's "statisticism" and reliance on aggregate-level statistics and Michell (2003) described the misleading notion of the "quantitative imperative" – the belief that all attributes of psychological phenomena are quantitative and therefore must be measured. Moreover, Freud observed in the early 20th century that psychologists tend to determine their research methods prior to defining their objects of study, implying

that statistical techniques typically drive the research questions proposed in the field (Robinson, 2001). Thus, it appears that psychology, in its pursuit to be accepted as a "real science," has relied heavily on quantitative methods and in turn has failed to appreciate the utility of qualitative methods. Interestingly, Latour (2000) argued that psychology, in fact, poorly imitates the physical sciences through its obsession with numbers, pointing out that many of the physical sciences rely heavily on qualitative description of phenomena.

Despite their lack of popularity in psychology, qualitative research methods have seen acceptance in many areas of the social sciences, particularly since the mid-twentieth century. Brinkmann et al. (2014) described the period from the 1960s and onwards as a type of "renaissance" for qualitative methods. In the second half of the twentieth century, many advocates of qualitative research began voicing their concerns with the limitations of quantitative methods. As a result, qualitative research gained momentum in the social sciences and has since become widely accepted in many fields. Meanwhile, qualitative methods are still often marginalized in modern psychological research. For example, in 2008, a petition with over 1000 signatures was presented to the American Psychological Association (APA) with the request to form a new division devoted to the use of qualitative methods (Gergen, 2018) The proposal was rejected by the APA, with some members arguing that qualitative methods are unscientific. In response, Gergen (2018) and others formed the Society for Qualitative Inquiry in Psychology (SQIP), which was later accepted as a sub-section of Division 5 of the APA. Nonetheless, the mandate of Division 5 remains heavily focused on evaluation and measurement (Lamiell, 2018) and SQIP appears to be a minor sub-section of the division.

5.2. The "Error" in Qualitative Research

In the same way that I aimed to illuminate and clarify the meaning(s) of error in quantitative research, I aim to do so for qualitative research traditions. Unlike quantitative research, however, the role of error in qualitative methods is not closely tied to historical developments in qualitative research practice. That is, "error" has not been a central concept to qualitative methodologies. Nonetheless, error and other concepts relevant to it have sometimes been discussed within the qualitative literature, albeit, "error" is oftentimes confused with notions of "validity" and "bias." That is, it is not always

clear whether qualitative researchers intend to use “error” in a way that is analogous to the uses in quantitative research. For example, Norris (1997) wrote,

> One practical way to think about the issue of validity is to focus on error and bias. Research whether quantitative or qualitative, experimental or naturalistic, is a human activity subject to the same kinds of failings as other human activities. Researchers are fallible. They make mistakes and get things wrong. There is no paradigm solution to the elimination of error and bias (p. 173).

Clearly, Norris (1997) used the concept “error” in an everyday sense to mean a “mistake.” This is quite different than its traditional usage in quantitative research where mistakes on the part of the researcher would not be considered error. Moreover, in quantitative research, error is a concept that has been relevant to the establishment of reliability. Although reliability is taken to be necessary for validity, it is not synonymous with validity. However, Norris (1997) appears to confuse the two.

One might question whether the concepts of “error” and “reliability” are relevant to qualitative methods. It might be argued that these are technical concepts germane to statistical methods and, as such, are not directly relevant to qualitative research practice. In particular, the objectivist view that underlies usages of the error concept within statistics is not as prevalent in qualitative research. In fact, qualitative research is often tied to philosophical views that emphasize interpretation and subjectivity. As mentioned before, there is diversity among various qualitative methods both in terms of their uses and in their associated philosophical foundations. However, as Brinkmann et al. (2014) noted, there are three general schools of thought that have been most influential for the domain of qualitative research. The first is the German tradition of hermeneutics which is based in the works of scholars such as Dilthey (2002), Gadamer (1960), and Heidegger (1927). Hermeneutics emphasizes the interpretive nature of texts as well as human life. From a hermeneutic perspective, psychological and social worlds are, by nature, interpretive, and humans are, by nature, self-interpreting creatures. The second school of thought that has been highly influential for qualitative research is phenomenology. This perspective originated with the works of Husserl (1954) in the early 20th century and was expanded upon by scholars such as Heidegger (1927) and Merleau-Ponty (1945). These thinkers promoted the idea that people are rooted in cultural, relational, and temporal environments. From this perspective, individuals make meaning of their lived experiences through engaging in an interpretative process that comes to constitute their

knowledge of the world. Accordingly, the first-person accounts of individuals are central to understanding psychological phenomena from a phenomenological perspective. Finally, the third philosophical tradition that has had a profound impact on qualitative research is that of American pragmatism associated with the works of scholars such as Dewey (1910) and James (1907). Pragmatists view "truth" as being embedded in human action and circumstance. As such, for a pragmatist, the goal of science is not necessarily to capture a fixed and objective reality but rather to understand reality in relation to one's actions and subjective experiences.

Collectively, the influential schools of thought described above acknowledge interpretation, subjectivity, and the circumstantial nature of knowledge. Given this, it is difficult to imagine how the quantitative "error" concept which is so deeply embedded in an objectivist view of the world could play a role in qualitative methods. However, it could be argued that like quantitative researchers, qualitative researchers also struggle with issues surrounding "truth" and what constitutes "truth." Although traditional quantitative and qualitative researchers might approach the question of truth from fundamentally different viewpoints, it seems that both classes of research methods will, at one point or another, come up against the issue of what counts as "truth." For example, many qualitative researchers have discussed the truthfulness of interview and participant reports when collecting data, as well as the value of a researcher being truthful about their own perspectives and biases regarding a research topic (e.g., Clark & Sharf, 2007; Flicker, 2004; Watt, 2007). Thus, there may be commonalities between the struggle for "accuracy" in quantitative and qualitative psychological research. Indeed, Wertz (1986) argued that both research approaches have attempted to deal with uncertainty in various ways and that both approaches assume and rely on some amount of variation in human phenomena.

In fact, the APA recently released reporting standards for qualitative and mixed methods (both quantitative and qualitative) research in which the term "reliability" appears in 3 instances. First, it is recommended that in their analysis section, researchers report a complete "description of coders or analysts and training, if not already described (interrater reliability, if used)" (Levitt et al., 2018, p. 39). Second, researchers are encouraged to "describe how issues of consistency were addressed with regard to the analytic process (e.g., analysts may use demonstrations of analyses to support consistency, describe their development of a stable perspective, interrater

reliability, consensus) or how inconsistencies were addressed" (Levitt et al., 2018, p. 39). Later in the document, recommendations are given for reports based on both quantitative and qualitative data. Here, it is stated that researchers should address the validity, reliability, and methodological integrity of the study. It is clarified, however, that the concepts validity and reliability are used here to refer to uses of quantitative data and the legitimacy of mixed methods. Thus, according to the APA's recently published standards, the concept of "reliability" is certainly relevant to qualitative research, although, it is acknowledged that the term "reliability" may hold different connotations for qualitative researchers than it does for quantitative researchers. Perhaps, then, features of research related to the quantitative notions of "reliability" and "validity" are discussed under different terms in qualitative methodology.

One concept often used in the context of qualitative methods that is related to the "legitimacy" or "quality" of research is that of "trustworthiness." Guba and Lincoln (1981) described four general aspects of qualitative research (or what they referred to as "naturalistic inquiry") that are addressed by the concept of trustworthiness. The first, *truth value*, is concerned with the level of confidence that a researcher has in the degree to which the results of a study are taken to be true. Second, *applicability* refers to how relevant a set of findings are to contexts outside of a current study. Third, *consistency* refers to whether the results of a study can be repeated if the study were to be re-run. Finally, *neutrality* is concerned with the degree to which the findings of a study have been biased by the perspectives and motivations of the researcher. Of these four topics, the third, consistency, was described by Guba (1981) as being related to the concept of reliability. Although Guba (1981) noted that reliability used here is not identical to its usage in the quantitative sense, he drew comparisons between quantitative and qualitative senses of "reliability."

> The naturalist is also concerned with consistency, and for the same reasons; naturalistic instruments no more than rationalistic ones are likely to yield credible (the analog of valid) results if they do not exhibit consistency. But consistency is a trickier concept for the naturalist than the rationalist. The latter, believing in a single reality upon which inquiry converges, can treat all instrumental shifts as error, but the naturalist, believing in a multiple reality and using humans as instruments – instruments that change not only because of "error" (e.g., fatigue) but because of evolving insights and sensitivities – must entertain the possibility that some portion of observed instability is "real" (p. 81).

Guba (1981) thus concluded that the term “consistency” has two meanings for qualitative or naturalistic researchers. It refers to “stability” under the quantitative sense of the term “reliability,” but it also refers to the degree of “trackability” in “explainable changes” when an instrument is employed (Guba, 1981, p. 81). That is, consistency under qualitative research also implies consistency in “trackable variance” that can be ascribed to sources other than random error (Guba, 1981, p. 81). Moreover, Guba argued that qualitative and quantitative inquiry differ in their relationships with the concept of “error.” Quantitative research aims to defend against or cover up error sources and qualitative research aims to “take account of the bewildering array of interlocking factor patterns that confront [researchers] and pose formidable problems of interpretation” (p. 84). Interestingly, Guba (1981) further provided solutions to the issue of consistency (which includes “reliability) for qualitative researchers that are extensions of strategies used in quantitative research. For example, a procedure described as “stepwise replication” was described by Guba (1981) as being “analogous to the “split-half” reliability of tests, in which two separate research teams… deal separately with data sources that have also been divided into halves” (p. 87).

The notion of “trustworthiness” has been widely adopted into the practices of qualitative researchers since it was first described by Guba and Lincoln in the early 1980s (Shenton, 2004). Thus, it appears that the concept of “error” may be relevant to qualitative research insofar as it is related to the notion of “reliability.” However, it is unclear what qualitative researchers might mean by “reliability” and if there is consensus within the domain of qualitative methods regarding the uses (if any) of “reliability” practices. As mentioned previously, Guba (1981) briefly discussed this issue; however, to achieve a more complete picture, it is necessary to explore how qualitative methods are discussed and used more broadly within contemporary psychology and the social sciences. In the next section, I outline the details of a project examining how qualitative researchers are currently discussing the notion of “reliability.” Specifically, I focus on recent journal articles and books that provide guidelines, recommendations, and summaries of how qualitative research is and/or ought to be implemented.

5.3. Reliability in Qualitative Research: A Content Analysis

The second aim of the overall current project is to examine whether the concept of “error” plays a role in qualitative research and what that role might be. In designing the

current analyses, I initially searched the qualitative research literature for pieces of methodological work discussing issues of error. However, my search yielded minimal results. It does not appear that the term error was being used broadly within the qualitative research community. Given this, I decided to instead examine uses of the term "reliability." As described in previous chapters, reliability is a concept that extends the ideas of error and true scores. Tests that have high reliability are believed to produce observations with lower associated amounts of measurement error. Thus, in the current study, I examine reliability in qualitative research as a proxy for the relevance of error for qualitative research methods. More specifically, I describe ways in which current qualitative researchers and methodologists in the social sciences are discussing the concept "reliability" in relation to uses of qualitative methods. Given that qualitative methods have not been widely accepted into psychology, the scope of the current analysis focuses more broadly on the social sciences in which a much larger literature on the uses of qualitative methods has been produced. The analyses described here are primarily descriptive and exploratory. Content analysis was employed to obtain a sense of both *how* the concept of "reliability" has been used and the *frequency* of these usages. In the concluding chapter, I take a more evaluative stance in which I compare uses of reliability in quantitative and qualitative methodology and implications for the concept of "error."

5.3.1. Method

5.3.1.1. Search Strategy

To obtain a sample of current readings discussing the relevance of reliability to qualitative research, I searched literature pertaining to qualitative research using two different methods on November 16th, 2017. First, I conducted a search through the *PsycINFO*© database in which I requested all published articles and non-published dissertations/abstracts from 2012-2017 that included the keywords "qualitative research" and "reliability" within the text. This search provided 125 results. Second, I conducted a library search through the Simon Fraser University library webpage of edited volumes, textbooks, and manuals published between 2012-2017 using the keywords "quantitative research methods" and "reliability." The word "methods" was added for this search because an initial search using only "qualitative research" resulted in thousands of

results, many of which were unrelated to the current project. Thus, the word “methods” was included to narrow the results for the current research focus. This second search provided a total of 172 results. In sum, I obtained an initial sample of 297 works.

The citations for each of these works were exported to an Excel file and a random ID number was generated for each citation. Next, citations were randomly chosen and scanned for relevance to the current project. A total of 60 works were randomly selected and scanned.[7] Of these 60 works selected, 34 were unrelated to methodology in qualitative research and/or did not use the word “reliability” in a way related to its uses in qualitative methods. Twenty-six of these works did consider reliability in the context of qualitative research and included discussions of how reliability might be used or considered in relation to qualitative methods. Given that all 26 works were published within the past 5 years, complete searchable versions were available online. This provided a useful tool for identifying parts of the works directly related to reliability. Thus, each of these 26 works was searched for the term “reliability.” All sentences and/or paragraphs related to the use of reliability in qualitative research were excerpted and pasted into a word document. This word document was then imported into NVivo Qualitative Data Analysis Software (2017) and examined using content analysis.

5.3.1.2. Analytic Strategy

Content analysis has typically been used as a general label denoting many different analytical procedures focused on identifying and sometimes quantifying common themes in qualitative data. Hsieh and Shannon (2005) outlined three different forms of content analysis based on its uses in the research literature. They described *directed* content analysis as a form of investigation that relies heavily on theory and previous research findings to inform the analytical process. *Summative* content analysis was described as more quantitative in the sense that it involves recording frequencies of uses of specific words or instances of specific content. On the other hand, *conventional* content analysis was described as much more exploratory and reliant on the data. In conventional content analysis, researchers examine the data at hand to identify common

[7] The sample size of 60 was chosen as a reasonable initial number of articles to review. My plan was, if after reviewing these 60 articles, saturation did not appear to be reached, I would return to the initial population of articles to randomly draw another subset. Saturation was reached; thus, I did not need to draw another sample.

categories of content. As such, this type of analysis aims for qualitative categories that "flow from the data" (Hsieh & Shannon, 2005, p. 1279). Given that my aim in the current study is to describe and explore ways in which reliability has been discussed in the qualitative methodological literature, I adopted the conventional form of content analysis.

First, data were initially read and re-read. Next, initial codes were created based on the data and excerpts were classified into one or more codes. These codes were then analyzed further and comments and further questions were assigned to each of the codes. Based on these comments, new aggregate categories were developed. The initial codes were combined into these new "superordinate" categories and descriptive labels were created as general categorical summaries. Initial codes were also retained in the form of "sub-categories." Finally, these superordinate categories, sub-categories, labels, and their associated frequencies are described and examined in the results section below.

5.3.1.3. Reliability of the Current Analyses

Given that the current content analysis focuses on the relevance of reliability for qualitative research, a related question concerns whether the reliability of the current analyses should be considered. I view the current analyses as being mainly qualitative and descriptive. Results are not necessarily meant to be generalized to the entire domain of qualitative research; rather, they are meant to invoke questions for further exploration and speculation. One might question whether my presentation of the articles in the sample is "authentic," "accurate," "trustworthy," or "reliable". To address this, I have chosen to follow a recommendation provided by Guba (1981). I have not applied the quantitative notion of reliability to the current analysis because it is not primarily quantitative or interpretive. Inter-rater reliability was not computed because the data reported were not rated or judged. Instead, I provide transparency of the results by including references to the 26 sources in the current sample in Appendix A. Readers are encouraged to directly consult these sources as a means of establishing "outsider" verification of the findings.

5.3.2. Results

Table 2 presents a list of the initial codes and their respective labels generated from excerpted works.

Table 2. Content Analysis Initial Codes

Code	Label	# of excerpts coded	# of sources coded	Category
Accuracy of Research	Mention of the "accuracy" of research or findings in the context of discussions of reliability.	2	2	a
Comparisons between quantitative and qualitative methods	Authors draw on similarities or differences between quantitative and qualitative methods in the context of discussions of reliability.	16	11	b
Emphasis on data and research questions	Discussion of how decisions to assess reliability should be based on researchers' data and research questions rather than on whether research is quantitative or qualitative.	1	1	b
Consideration of epistemology	Discussion of epistemology in the context of reliability.	3	3	a
Consideration of community	Assessing whether research addresses the community rather than focusing on reliability.	1	1	c
Reliability as "quality"	Reliability discussed in the context of assessing the quality of research.	2	2	d
Inter-rater reliability	Reliability discussed in the context of achieving agreement amongst raters.	5	5	e
Reliability irrelevant for qualitative research	Authors imply and/or discuss whether reliability is irrelevant for qualitative research.	7	7	b
Reliability determined by evaluations from "others"	Individuals besides the primary researchers of the project help to establish the reliability or quality of the project.	3	3	c
Qualitative research examines variability, quantitative research eliminates it	Reliability is relevant to quantitative research because of the aim of eliminating variability. Qualitative research does not share this aim.	1	1	b
Qualitative terms used to replace quantitative terms associated with reliability	Authors discuss terms in the qualitative research literature (e.g., trustworthiness) that are deemed as alternatives to the quantitative notion of reliability.	8	7	b & d

Code	Label	# of excerpts coded	# of sources coded	Category
Emphasis on the relationship between participant and researcher	The relationship between participant and researcher described as important for reliability of qualitative studies.	1	1	c
Reliability consists of the coherence of measurements	Reliability described as consisting of the coherence of measurement instruments.	1	1	d
Reliability defined as consistency	Reliability described as the level of consistency in observations/measurements/findings.	9	6	d
Reliability defined as repeatability of observations	Reliability described as the repeatability of observations/findings.	4	3	d
Difficulty in establishing reliability a problem for qualitative research	Authors discuss difficulties/issues in thinking about and using reliability in the context of qualitative research. They describe these difficulties as problematic for qualitative methods.	6	3	b
Reliability as a positivist concept	Reliability is associated with positivism and described as stemming from positivist ideals.	9	8	a
Reliability as "rigor"	Reliability defined as "rigor" for qualitative research.	4	3	d
Reliability an important concept for qualitative research	Reliability described as being important and useful for qualitative research methods.	4	4	b & f
Reliability discussed in the interview context	Establishing reliability of interviews conducted with participants.	6	2	e
Reliability includes interpersonal replicability	Reliability includes interpersonal replicability.	1	1	c
Reliability increases credibility of research	Assessing reliability increases the credibility of a study.	2	2	f
Reliability stems from an objectivist viewpoint	Reliability discussed as stemming from a viewpoint that adheres to objectivism.	4	4	a

Code	Label	# of excerpts coded	# of sources coded	Category
Reliability through consensus	Reliability achieved in qualitative research through consensus.	4	2	c
Reliability through examination of patterns of thought and behaviour	Authors discussed the establishment of reliability in qualitative research through the examination of thought and behaviour.	1	1	c
Reliability through reflexivity	Reflexivity discussed in the context of qualitative reliability.	2	2	c
Reliability through transparency	Researchers are expected to be transparent in the research process. This is discussed in the context of reliability.	1	1	c
Reliability through triangulation	Triangulation discussed as a method for obtaining reliability.	5	5	c
Reliable researcher	Reliability discussed in the context of a "reliable researcher."	1	1	c
Consideration of role of researcher*	The role of the researcher must be considered when determining reliability.	3	3	c
Subjectivity in qualitative research	Subjectivity discussed in the context of reliability.	2	2	a
The "truth" in qualitative research	References to "truth" in the context of reliability.	4	2	a
Validity prioritized over reliability	Validity discussed as being more immediately important than reliability for qualitative research.	6	5	b
Verification	Reliability addressed through "verification."	2	2	d

5.3.2.1. Content Categories

Based on comments that were developed from these initial codes, 6 superordinate categories of codes were created and each of the initial codes were placed into a category (2 of the codes were placed into 2 different categories, these are identified in the table above). Table 3 presents these 6 categories along with their respective descriptions. A column presenting frequencies of the number of instances coded within each category is also provided. Given that the coding categories were developed based on initial codes, it is possible that the same source was coded into the same category multiple times because it was included in more than one initial code

within that category. Thus, information regarding number of instances coded, rather than number of sources coded, is provided. However, in the following discussion, categories and their respective codes are described in conjunction with the number of sources assigned to each code.

Table 3. Content Analysis Superordinate Categories

	Category	Description	# of instances coded
a	*Philosophical considerations*	This category consists of excerpts that discussed the philosophical underpinnings of "reliability." Most commonly, this is associated with positivism and the quantitative research tradition. The notion that reliability is tied to an objectivist view and that qualitative research is based in subjectivity is also discussed.	17
b	*Quantitative vs. qualitative*	In this category, the issue of whether reliability is an appropriate concept for qualitative research is discussed. Comparisons are also made between quantitative and qualitative research methods.	39
c	*Criteria for assessing reliability*	This category includes excerpts that discussed various criteria and/or ways to judge the reliability/quality/rigor of qualitative research.	21
d	*Defining reliability*	Reliability was defined in a variety of different ways in the sample of excerpts. Some authors described different terms that would replace the concept of "reliability" for qualitative research, while others described what reliability refers to for qualitative research.	24
e	*Using reliability*	Several sources described the context in which reliability would be examined a qualitative research study.	7
f	*Benefits of considering reliability*	This category includes excerpts that described some of the advantages of assessing and/or having a notion of reliability for qualitative research.	6

Philosophical Considerations

Three sources in the current study mentioned epistemology in the context of reliability for qualitative research. These authors stated that qualitative research of high "quality" should ask epistemological questions and be "epistemologically sound." One of the sources argues that reliability is only relevant for qualitative methods insofar as a quantitative epistemological outlook is adopted within the qualitative research. Indeed, reliability was often aligned with a certain perspective on knowledge and truth. In particular, 8 sources associated reliability with positivist philosophical groundings. These

sources described positivism and postpositivism as being the dominant frameworks in psychology and the philosophical underpinning of reliability and quantitative research. Based on these arguments, some authors described reliability as being suitable to a positivist worldview which was often juxtaposed with the constructivist worldview of qualitative researchers. Moreover, 4 sources tied reliability to an objectivist view; however, only one of these sources provided some definition of what was meant by objectivity, explaining that reliability assumes the pursuit of an objective "truth." One other source besides this one also mentioned "truth" in the context of reliability, explaining that verification of truth is difficult to achieve in qualitative research. In the same vein, 2 sources mentioned subjectivity in the context of reliability. One of these sources argued that establishing reliability in the context of inter-rater agreement is problematic because raters might hold a shared, yet subjective and interpretive, perspective. Similarly, another source mentioned that reliability in qualitative research represents a shared perspective that is temporal and situational rather than an ultimate and objectivist truth. Lastly, 2 sources mentioned the term "accuracy" when discussing reliability issues in qualitative research. One of these articles questioned whether accuracy was an appropriate aim for qualitative research, while the other described the difficulties in achieving accuracy with qualitative interview data.

Quantitative vs. Qualitative

The largest proportion of codes in the current analyses fell under the initial coding category of "comparisons between quantitative and qualitative methods." These included excerpts that described reliability in the context of comparing quantitative and qualitative research. Specifically, 11 of the sources in the current study described reliability in this context. A common sentiment within these 11 sources was the notion that reliability stems from quantitative research practices which are positivist by nature and which can be contrasted with the constructivist tradition of qualitative research. Two excerpts also mentioned that some qualitative researchers in disciplines such as the health sciences might conduct reliability analyses because they feel pressured to meet positivist expectations. One source argued that quantitative research aims to ignore and/or eliminate variation between and within individuals, while qualitative research examines such variation. Indeed, a total of 7 sources in our sample mentioned that reliability analysis might not be as relevant for qualitative research as it is for quantitative. Interestingly, 5 of the sources in the current sample described validity as

being prioritized over reliability in qualitative research and several of these authors stated that reliability is not as important for establishing validity in qualitative research as it is for quantitative research.

Other sources indicated that all research, whether quantitative or qualitative, takes more than just reliability and validity analyses to be considered "good" research while others stated that reliability is equally as important for qualitative research as quantitative research, although there is less methodological literature on the uses of reliability in qualitative methods. Two sources acknowledged that reliability plays a role in qualitative research, although in a different way than in quantitative research. For example, one source argued that the term "verification" is more suitable to qualitative research. This is in line with another common theme within the category of comparing quantitative and qualitative methods which involved using new terms to describe the quality of qualitative research rather than the "quantitative" terms of reliability and validity. Seven sources from our total sample mentioned alternative terms deemed more suitable for assessing the quality of qualitative research. These included terms such as: trustworthiness, verification, rigor, credibility, dependability, confirmability, reflexivity, and consistency. However, 3 sources acknowledged that assessment of "reliability," although perhaps useful for qualitative research, is more difficult to establish than in quantitative research. This was described as being due to a number of reasons, including the presence of the observer in the research context, small sample sizes, and lack of efficient strategies for establishing the quality of qualitative research. In addition, 4 sources in total acknowledged that reliability could be a useful concept for qualitative research. One of these authors stated that qualitative researchers face many of the same challenges as quantitative researchers, while another argued that qualitative researchers also need to provide adequate evidence to support the soundness of their conclusions. Finally, 1 source questioned the legitimacy of a quantitative-qualitative dichotomy and argued that the question of whether reliability is relevant to a research project depends on the aims and data collected.

Criteria for Assessing Reliability

Each of the different sources in the current sample emphasized varying criteria for establishing reliability (or its qualitative "alternative") for qualitative research. Single sources recommended each of consideration of the role of the project within a

community, the relationship between the participant and the observer, the transparency of the researcher and results, the reliability of the researcher, consideration of patterns in the thoughts and behaviours of participants, and whether a study had “interpersonal replicability.” Three sources mentioned that qualitative studies should consider the role of the researcher in the observation process. In addition, 3 sources mentioned that reliability for qualitative research can be assessed by “others” with outside perspectives on the study at hand. Two sources described reflexivity as a process for establishing reliability. This involves personal reflection by the researcher regarding the research process and his or her biases. Other criteria and methods for assessing reliability included procedures for reaching consensus (particularly in the context of inter-rater agreement), triangulation (i.e., verifying findings through different perspectives and procedures), and verifying findings through member checking.

Defining Reliability

Many sources in the current sample used terms that they described as “alternatives” to the terms “reliability” and “validity” in quantitative research. Overall, 7 sources argued for replacing these terms within the qualitative literature and offered alternative such as “trustworthiness,” “dependability,” and “confirmability.” More specifically, 2 sources defined reliability under the broader umbrella term of “quality” while 3 sources argued for the use of the term “rigor” and 2 sources for the term “verification” rather than the terms “reliability” and “validity” (i.e., these terms would encompass aspects of both “reliability” and “validity.”) Several sources attempted to stay close to the definition of reliability in quantitative research and thus 6 sources described reliability as “consistency” and 2 as “repeatability” of observations. Lastly, 1 source described reliability as the “coherence” of observations.

Using Reliability

The most common scenarios in which the use of reliability was described in the current sample of sources were within interview contexts and the process of establishing inter-rater agreement. Several other sources also considered reliability of narratives, observations, and diaries. Two of the sources in the current sample explicitly discussed the relevance of reliability for interview research. For example, one source discussed how qualitative interviews are more flexible and less constrained than quantitative interview structures. An implication of this, they argued, was that reliability would be

decreased; however, such interviews would likely have better face validity. Another source discussed issues with verifying the "truth" in the context of interview studies, stating that one way to mitigate this problem is to return to the source and ask participants to "check" the interpretation of the results. Five of the sources in the current study discussed the role of reliability when assessing agreement between2 or more coders and/or raters. Recommendations from these authors regarding how to establish reliability in this context varied. Suggestions included using multiple coders, providing statistical indices of inter-rater agreement, aiming to reach consensus over statistical reliability, and providing explicit information regarding coding procedures.

Benefits of Considering Reliability

As a final category, I considered sources that explicitly stated that assessing reliability in qualitative research is advantageous or should be carried out. Five sources fell into this category. One source stated that reliability must be addressed in qualitative research in order to alleviate similar problems that are faced in quantitative research. Another source drew attention to the issue of "measurement" instruments in the context of qualitative research in which raters numerically code qualitative data (i.e., in this case, the "measurement" instruments would be akin to the rating guidelines/manual). This source noted that whether such instruments provide results that are consistent has been a point of contention in qualitative research. A third source argued that reliability should be a focus of all researchers and that it is important for qualitative researchers to establish the reliability of their findings. The final 2 sources emphasized the importance of assessing reliability in qualitative research by arguing that strong reliability improves the credibility of qualitative studies.

5.3.3. Discussion

The aim of the present content analysis was to explore ways in which current authors of qualitative research methodology discuss the concept of "reliability" in relation to qualitative research. This was carried out as part of the larger goal of identifying whether the concept of "error" is relevant to qualitative research and, if yes, how so? Thus, reliability was used as a proxy for examining the role of error in qualitative research. The results of the initial search for sources revealed that although the term "reliability" appeared frequently within the qualitative methods literature, many

references to the term were unrelated to its usage in research (i.e., of the 60 sources randomly selected for review, only 26 (43%) were related to the present research aims). Although inconclusive, this might suggest that qualitative methodologists are, in general, not very concerned with the concept of "reliability" for qualitative research. However, as previously mentioned, establishing "trustworthiness" is an important aspect of qualitative research; thus, it may be the case that concepts related to the quantitative notion of reliability are being discussed in the qualitative methods literature under terms other than "reliability." Indeed, one of the resulting superordinate categories of the current content analysis pertained to authors defining and describing reliability and related concepts in various ways. Seven of the articles from the sample (27%) replaced the quantitative notion of reliability with a different term that was meant to capture reliability (and sometimes other aspects of "quality") for qualitative research. Terms such as "dependability" and "consistency" were used by these authors to represent the level of reliability of a given qualitative research project. These same terms might be used in the context of quantitative research where reliability is more often defined in terms of the "repeatability" of results. In this way, one can see some overlap in the uses of the term in both research traditions. In quantitative research, the 2 core tenants of reliability and validity often speak to the level of "rigor" of a given research project.[8] Likewise, multiple qualitative sources in the current study suggested that the term "reliability" be either replaced or encapsulated by the term "rigor." However, what is meant by the term rigor may differ between quantitative and qualitative traditions as well as within. For example, in qualitative research, a rigorous project is one that takes into consideration researcher bias and the situational context, two factors that are less often considered in quantitative research. One reason for differences in terminology may be differences in aims of quantitative and qualitative researchers. Silverman (1993) argued that qualitative researchers strive for authenticity rather than reliability of results. However, it is likely that most quantitative researchers would also argue that they strive for authenticity in their research. Thus, perhaps quantitative and qualitative researchers share some common aims. I will re-visit this point later in this section and in the following chapter.

[8] Note, however, that reliability and validity are not viewed as being central to the "rigor" of a study by many psychometricians. That is, quantitative researchers in psychology often hold these two concepts up to a standard for which psychometricians did not initially intend them. For example, the concept of reliability has very little connection to the overall rigor of a quantitative research project (i.e., it does not guarantee reliability of your theory, design, etc.).

One stark, but unsurprising, difference between the ways in which reliability (and terms related to it) are described in qualitative and quantitative methodology is the lack of mathematical and technical definitions in qualitative research. On one hand, this is completely reasonable given that qualitative research is, by definition, not based on quantitative or mathematical foundations. However, what's interesting is that the concept of "reliability" in psychological research is rooted in psychometric traditions and is thus perhaps *best* defined mathematically in quantitative research. In fact, one might argue that reliability *is* a mathematical and highly technical concept. That is, reliability in quantitative research is defined computationally as a ratio of true score variance to observed score variance. In the current sample, perhaps unsurprisingly, very few sources discussed reliability computationally (i.e., few sources made mention of *calculating* reliability). This was usually done in the context of mixed methods research and/or establishment of inter-rater agreement. However, by far, when qualitative methodologists in our sample discussed the relevance of reliability for qualitative research they did so in non-technical and non-computational ways. It appears that whereas quantitative researchers aim to use reliability in a very technical sense, qualitative researchers refer to the term in more of an everyday sense.

Reliability is defined in its ordinary sense by the online English Oxford Dictionary as "the quality of being trustworthy or of performing consistently well" (Oxford University Press, 2018). The terms "trustworthy" and "consistency" are two that appear quite often in the qualitative methodological literature in relation to the term reliability. Thus, it appears that qualitative researchers are indeed aiming to use the term in primarily an ordinary rather than technical or computational sense. If this is the case, then should qualitative researchers feel obligated to even refer to the term "reliability?" Do such references invoke the quantitative tradition of the term? As mentioned, reliability in psychometrics is first and foremost defined computationally as a ratio of variances. Variances are average squared deviations from a central value. As such, the "central" value plays an important role for the meaning of reliability in quantitative research. As described in chapters 1-4, the central value, which is typically an average or mean score, has been interpreted in various ways throughout the history of statistics and psychology. Early uses of statistics in the field of astronomy viewed the central value conceptually as the "true" value and an arithmetic mean was calculated as an estimate of the "truth." As previously described, varying interpretations of this "true" value were

proposed over time; however, the notion of truth in quantitative methodology has commonly been viewed through an “objectivist” lens. That is, quantitative methods are generally practiced from a perspective according to which the objects of study exist independently from the observer. The implication for the quantitative notion of reliability, particularly regarding measurement, is that the true existence of the object of study does not necessarily change with repeated measurements. Thus, the notion of reliability as a way of capturing consistency over repeated measurements to obtain an “accurate” estimate of a true value is valid from an objectivist point of view. As described in chapter 4, the goal of “accuracy” is a theme that has remained mostly constant throughout the history of statistics and quantitative psychology.

Four of the 26 sources (15%) in the current sample tied reliability to objectivism while 2 sources (8%) mentioned that qualitative research is founded in subjectivist views. From a subjectivist perspective, the “objects” of study in psychological and social research do not exist independent from their context and are inseparable from the observer. Given that many qualitative researchers ascribe to a subjectivist way of thinking about their objects of study, how does a concept such as reliability, which is founded in objectivist views of the world, fit into qualitative methodology? For many of the authors in the current sample, the answer is quite simple: it doesn’t. Seven sources (27%) questioned and/or denied the relevance of a quantitative notion of reliability for qualitative research. Authors questioned whether a concept that stems from a tradition founded in objectivist beliefs is relevant for a domain of research that openly accepts subjectivism. Indeed, 2 sources (8%) from the current sample questioned whether an aim of accuracy was relevant or even plausible for qualitative research.

Three sources (25%) in the current sample also stated that researchers should consider the epistemological stance from which they are working under when determining if reliability is relevant for their research. It was argued that a constructivist view of epistemology sees no divide between knowledge and knower and, therefore, the concept of reliability is only relevant for those qualitative researchers adopting a positivist framework. Indeed, Madill, Jordan, and Shirley (2000) argued that qualitative researchers need only concern themselves with adopting criteria for the quality of research from the quantitative sciences *if* they approach their research from a realist perspective. Yet, rather than realism, positivism was cited as being the founding philosophy behind the concept of reliability by 8 sources (31%) in the current sample and

it was also the *only* philosophical framework associated with reliability (post-positivism was also mentioned by one of the sources). Authors stated that reliability was a "positivist" standard and a "positivist" term. However, none of the sources explained *why* reliability would be tied to positivism, nor did they explain what was *meant* by the term positivism; although, several described quantitative research methods as stemming from a positivist framework. Moreover, it should be noted that some level of "realism" is inherent in almost all philosophical viewpoints, including constructivism (Dreyfus & Taylor, 2015). Realism does not necessarily imply an objectivist perspective.

The association between positivism and reliability is noteworthy given that reliability in psychology is a concept that stems from the early works of psychometricians who initially developed tools such as factor analysis to investigate hidden underlying causes of mental phenomena. It is a central tenet of positivism that science can only be conducted using tools of direct observation. Hidden underlying causes are not amenable to the practices of science, and thus, a positivist would likely not approve of the use of reliability in the study of such unobservable attributes. As mentioned in previous chapters, Pearson, who helped develop correlation analysis mathematically, was a strong positivist who rejected the idea of "uncovering" unobservable causes (Gigerenzer, 1987). Thus, Spearman (1904a; 1904b) and other early psychometricians who would later adopt correlational techniques in the pursuit of measuring "unobservables" would do so in opposition to some traditional positivists' views. Although Comte (1975), the founder of positivism, believed that only objective knowledge is relevant in science, he did not believe that all knowledge is necessarily objective. The study of unobservable causes would fall in the realm of subjective knowledge, which Comte would likely argue should not be studied using quantitative science.

Throughout the sources in the current sample a common theme in discussing the relevance of reliability for qualitative research was the comparison of quantitative to qualitative methods (11 sources (42%) explicitly made these comparisons). Authors described quantitative research as being aligned with positivism and qualitative research as being aligned with constructivism. Some authors also discussed the pressure that qualitative researchers in quantitative-dominant fields feel to use quantitative principles such as reliability in their work regardless of whether they are fully relevant to their research methods. Michell (2003) explored the dichotomization of quantitative and qualitative methods and the widely-cited notion in qualitative research that quantitative

methods are founded in a positivist framework that is not welcoming to qualitative methods. According to Michell (2003), positivism is not inherently tied to quantitative methods and instead encourages uses of diverse methodologies in the study of psychological phenomena. Thus, although reliability may be primarily defined computationally (i.e., quantitatively) and stem from objectivist ideas about knowledge, it is not inherently tied to a positivist framework, nor does a positivist framework always imply the use of quantitative methods.

Nonetheless, it appears that qualitative researchers want to be able to talk about the trustworthiness of their research while also acknowledging the subjectivity of knowledge. As such, qualitative researchers from the current sample discussed various ways of addressing the reliability of a study that are quite different from calculating a quantitative index of reliability. They mentioned ways of establishing the trustworthiness of a study that acknowledged the relationship between knowledge and knowers. For example, authors suggested verifying how a study reflects the community within which data was collected, examining observer bias and the relationship between researchers and participants, using observer personal reflections, and considering "interpersonal reliability." Although some authors did discuss the uses of inter-agreement indices, it was clear from the current data that reliability for qualitative research goes beyond quantitative indices and should approach knowledge production through a subjective lens.

Based on the current sample, one might conclude that qualitative researchers often care about issues related to reliability; however, how they define reliability is much more consistent with everyday uses of the term rather than technical computational uses. Perhaps, then, qualitative researchers need not worry about meeting "quantitative" expectations and should forego the notion of "reliability" for qualitative research entirely. Instead, they might opt to adopt other terms that better suit their needs. Indeed, a question for future explorations in this area might be to examine whether issues pertaining to "reliability," and consequently, "error," might be discussed in the qualitative methodological literature under different terms. Based on my current analyses, terms that might be of interest include, "trustworthiness," "credibility," and "dependability." As a first step to this end, in the following sub-section I describe an initial search and preliminary analyses of qualitative methodological works that refer to the term "trustworthiness."

5.3.3.1. A Brief Exploration of Trustworthiness

As with my exploration of reliability, I limited my search of sources pertaining to trustworthiness to publications within the past ~5 years (2012 – April 2018). Searches were conducted through the Simon Fraser University (SFU) library books catalogue and the *PsycINFO*© online database. The *PsycINFO*© search consisted of the keywords "trustworthiness" *and* "qualitative research." Here, the "and" signifies that sources must consist of both sets of keywords. The SFU library search consisted of the keywords "trustworthiness" *and* "qualitative research methods." The word "methods" was added to this search to once again narrow the focus of the search to methodology. My search of journals within the *PsycINFO*© database provided 94 results, while my search of books within the SFU library catalogue resulted in 516 sources. Using the strategy previously described, a small preliminary sample of 20 sources from these searches was randomly selected for review. Sources were excluded if they did not discuss the implications of trustworthiness for methodology in qualitative research but rather only stated that they assessed trustworthiness in an empirical study. Of the selected 20 sources, 12 were related to the topic of trustworthiness and methodology in qualitative research. The term "trustworthiness" was searched in each of the source documents and excerpts were uploaded into Nvivo Qualitative Data Analysis Software (2017) for analysis.

I conducted an initial read of each of the extracted excerpts for exploratory purposes. Although a complete formal analysis was not undertaken, several preliminary codes were apparent. Trustworthiness was described as either being equivalent to validity, or being a form of validation, by 6 sources. In addition, 2 sources mentioned that "trustworthiness" is a more appropriate term for qualitative research than the terms "reliability" and "validity." One of these sources further tied the terms "reliability" and "validity" to positivist views. Moreover, 3 sources described trustworthiness as a means of gauging how "accurately" a study captures experience. Several different methods of ensuring trustworthiness were described, including member checks (checking the results of a study with the participants; 3 sources), using reflexivity (researchers being open about their subjective biases; 2 sources), and providing transparency and evidence of results (2 sources).

One speculative theme that arises from this preliminary exploration is the multiple references to trustworthiness as being akin to a form of validity. In my analyses of uses of the term “reliability,” I found a few authors mentioned that validity is far more important for qualitative research than reliability. It appears that what might be considered “trustworthy” for a qualitative researcher is related to how “valid” a study’s results are. However, what is meant by “valid?” Surprisingly, the word “accurate” appeared in 3 of the sources I examined. Authors described how a trustworthy qualitative study was one that accurately portrayed or “represented” the lived experiences of participants. To assess this level of accuracy, techniques such as member check, transparency, and reflexivity were suggested. The mention of accuracy and the use of member checks to ensure that interpretations of qualitative data are representative points to the notion that the aims of qualitative researchers may, in some ways, be quite similar to those of quantitative researchers. Indeed, the term “trustworthiness” would be a beneficial concept to further explore within both qualitative and quantitative research methodologies. Comparisons between the goals of quantitative and qualitative research will be discussed in the final chapter.

In sum, it appears that reliability, or some version of it, is of interest to the sample of authors in the current study. In general, there is a sense that the quantitative notion of reliability is too “objective” for qualitative researchers adopting a perspective that emphasizes a subjectivist epistemology. At the same time, it appears that qualitative researchers really want to use the notion of “reliability” in an everyday sense rather than in its technical sense tied to quantitative science. However, there were some authors in the current sample who believed that the struggles of quantitative and qualitative researchers are not very different and 2 authors argued that assessing reliability increases the credibility of qualitative research. To be fair, these advantages of reliability were mainly considered in the context of discussing measurement instruments for qualitative research and, thus, these authors are likely considering scenarios in which qualitative researchers are using rating methods.

However, it is worth considering whether quantitative and qualitative researchers have shared aims. Despite differences in how quantitative and qualitative researchers might define the concepts of truth and knowledge, and the relationships between people, truth, and knowledge, *all* kinds of researchers share the goal of making justifiable and credible claims. Moreover, my preliminary analyses of the term “trustworthiness” indicate

that “accuracy” may also be a concern for qualitative researchers. According to Wertz (1986), both quantitative and qualitative research domains have, at some level, acknowledged issues of uncertainty and have strived to achieve greater certainty. That is, all researchers strive to be as certain as possible about the results of their research. Although interpretation and multiple perspectives are acknowledged and embraced in qualitative studies, qualitative researchers have also discussed methods for ensuring “authenticity,” “trustworthiness,” and “credibility.” In addition, it is possible to misrepresent participant’ ideas, thoughts, and actions in qualitative research. For example, Borland (1991) tackled the issue of interpretive conflict, noting that it is possible for a researcher’s personal bias to cloud the authenticity of qualitative research, implying that there is some level of “accurate” representation that should be considered when conducting qualitative research. Moreover, although it appears that qualitative researchers more openly accept subjectivity than quantitative researchers, Wertz (1986) pointed out that objectivity and subjectivity are not completely independent and that the former is always open, to some extent, to the latter. In the final chapter, I re-visit conclusions from my examinations of the quantitative history of error and uses of reliability in qualitative research. In doing so, I consider further whether quantitative and qualitative research have shared goals and perhaps even shared interpretations of “error.” I also consider the implications of the current analyses for the field of psychology.

Chapter 6.

Quantitative and Qualitative Approaches to Error: Implications for Psychology

Thus far, I have addressed two research questions initially outlined in Chapter 1. The first concerned examining the history of the "error" concept in statistics and, more importantly, quantitative psychological research to clarify the meaning(s) of error in psychology. The second concerned the relevance of "error," by proxy of the concept of "reliability," to qualitative research methodology. In this final chapter, I address my third and final research question. Specifically, I draw on my findings in relation to my first 2 research questions to make comparisons between the role of "error" in quantitative and qualitative research. I then re-visit all three of my research questions and discuss implications for the field of psychology.

6.1. Comparing "Error" under Quantitative and Qualitative Methods

6.1.1. "Error" in Quantitative Methods

Based on the historical review and analysis conducted in the first part of this project, it is clear that the notion of "error" in quantitative methods stems from an objectivist view. Astronomers believed that their objects of study existed independent of observer influence. The error concept was used to denote the difference between an observer's measurement of some property of an object (or of relations between objects) and the property itself. Although measurements and the errors associated with them could vary over repeated observations (i.e., they could be represented with random variables), the true existence of the object under study was taken to be fixed. As such, scientists strived for *accuracy* of observations, which could be achieved through the minimization of observational errors. Such errors were conceived of being due to random and uncontrollable forces rather than to any influence on the part of the observer, hence, the objectivist foundational viewpoint from which the random variable model stems. This objectivist interpretation of error as the difference between true values and observations, along with an emphasis on the accuracy of scientific pursuits,

was carried forward in 19th century studies of psychophysical, human, and mental phenomena.

However, when adopted within psychological and social domains, the concept of error became complicated due to the increased complexity of the objects that were under study. Studies of mental, sensory, and behavioral phenomena were typically conducted with repeated observations of the same object and/or event under *varying* conditions. For example, Wundt often observed the same behavior in the same individual but across varying time points. On the other hand, Galton observed the same phenomenon across varying people. Although these studies were conducted at different levels (i.e., intra-individual vs. inter-individual), in both cases, varying conditions were not solely the product of circumstantial or environmental differences; rather, they were also due to the dynamic nature of the objects of study themselves (i.e., people).

In the latter half of the 19th century, scholars took advantage of the "error" observed in research involving people. For example, Galton interpreted "error" at the inter-individual level (between individuals) as variation. As a result, he spearheaded one of the most important statistical techniques of the 20th century, correlation analysis, to examine such variation. These techniques were later adopted by 20th century psychometricians as a means of validating psychological tests meant to "tap into" what was interpreted as being sources of psychological phenomena. Psychometricians argued that the study of psychological phenomena carried greater amounts of random error than studies in the physical sciences, and therefore, a psychometric theory was necessary for psychological sciences. There is also an underlying objectivist framework that is foundational to psychometric theory. Namely, the conception that psychological objects of study exist as "unobservable" attributes independent of observer influence but that these psychological objects can be indirectly measured through associations between manifest variables (i.e., the associations are taken to be due only to the causal forces of "unobservable" phenomena).

Finally, as discussed in Chapter 4, many psychometricians opted to focus on the operational uses of concepts such as "error" and "true scores" in their writings rather than on related ontological or epistemological issues. Thus, it is my impression that the natures of psychological phenomena are often not addressed in quantitative psychological research. Instead, it appears that the field has implicitly adopted an

objectivist view while focusing primarily on the operational uses of statistical tools rather than their conceptual implications. Thus, my exploration of "error" under quantitative methods has led to the following key points:

- The notion of "error" in statistics stems from an underlying objectivist viewpoint in which observers are independent from that which is observed.
- "Error" was a primary focus of psychologists and, particularly, psychometricians in the 20th century, who viewed psychology as having to contend with more random errors than the physical sciences. For this reason, it was argued that psychology needed a sound psychometric theory.
- The notion of "error" has been used to describe variation at both intra-individual and inter-individual levels; however, the distinction between the two is not always made clear.
- A persistent and underlying goal of most, if not all, quantitative research in psychology is *accuracy* of observations.
- Psychometricians often focused on operational uses of the concepts "true score" and "error" rather than on their philosophical implications.
- Degrees of accuracy and certainty are captured mathematically (e.g., through the concepts of reliability and standard error of measurement) in quantitative methods.

In consideration of these points, it is important to note that it would be quite rare to find a psychologist who would deny the subjectivity inherent in the study of people. Indeed, it is the variation in traits and behaviors between and within individuals that makes the study of psychology interesting and even feasible. Perhaps then, there is room to talk about subjectivity in quantitative methods. However, what would be the implications of this for uses of the "error" concept? An exploration of a methodological domain that openly addresses the subjectivity of psychological phenomena (i.e., qualitative methods) may help to resolve this question.

6.1.2. "Error" in Qualitative Methods

The content analysis portion of the current project examined the relevance of "error" for qualitative research methods by investigating uses of the term "reliability" in methodological qualitative papers. Given that the term "error" does not appear often in the qualitative methods literature, reliability was used as a proxy for assessing whether error plays a role in qualitative research. In quantitative research, reliability is taken to

represent the consistency or repeatability of measurements. A high index of reliability indicates less error associated with such measurements. Moreover, the concept of reliability stems from the random variable model conception of error being the deviation between observed values and a true fixed score. Therefore, reliability is inherently tied to the notion of error in quantitative methods and to an objectivist view of objects under study. Indeed, some authors in my sample of qualitative methodological works declared that assessment of reliability is only relevant for qualitative methods if the researcher is adopting a view that ascribes to an objectivist viewpoint. In these situations, researchers would presumably prioritize the accuracy of their results, and therefore, an assessment of reliability would be appropriate.

At the same time, there were several authors in the sample that discussed the issue of feeling compelled to conduct reliability analyses solely because it was an expectation of their quantitative-dominant field. This speaks directly to a systemic and cultural issue within academic disciplines. It also speaks indirectly to the fact that some qualitative researchers would feel no need to address reliability if it were not for the conventions in their field. There is, therefore, a political and social dimension to uses of concepts tied to "error" within research domains that should also be taken into consideration. I re-visit this issue in the final "implications" section.

Within the current sample of papers examined, it also appeared that qualitative methodologists are using replacement terms to convey similar meanings to that of the term reliability. This includes terms such as dependability, confirmability, and trustworthiness. Each of these terms potentially speaks to the consistency or repeatability of an observation. To obtain a sense of what might be meant by these replacement terms, I conducted an initial exploration of uses of the term "trustworthiness" in the qualitative methodological literature. I found that qualitative researchers were in fact using the term in a sense that implies approximation to truth. That is, several authors described trustworthiness as having implications for the representativeness and accuracy of qualitative research results. This implies that one goal of qualitative research might be to accurately portray objects under study. In this sense, it appears that qualitative researchers are interested in achieving a sense of "truth" and that there perhaps is room for "error" when pursuing such "truth." However, what is meant here by the terms "truth" and "accuracy" may vary.

Indeed, it was mentioned several times in the qualitative methodological papers examined that qualitative researchers ascribing to a constructivist, interpretivist, or phenomenological perspective on knowledge and reality view truth as being multi-faceted, situational, and contextual. Thus, an "error" concept for qualitative methods might also be much more multi-faceted than error as invoked in quantitative research. In addition, given that qualitative researchers do not often make use of numeric representations, "error" in qualitative terms would likely not be quantifiable. However, would qualitative researchers ascribe to a notion of "error" in which the concept represents differences between observations and truth? To do so, would one need to accept the idea that a person's observations are independent of truth? Alternatively, could the concept of "error" be adapted to suit a view that accepts greater subjectivity in truth (i.e., where truth is not fixed and independent of the observer). While I did find mentions of accuracy in my search of sources discussing "trustworthiness," I found 2 sources questioning whether accuracy is a legitimate aim for qualitative research in my search of sources that mentioned "reliability."

Finally, in the same way that authors used alternative terms to describe "reliability," it may be the case that "error" is also being discussed in qualitative methodological literature under alternative terms. Recall, that within quantitative methods, random error is distinguished from systematic error and further from any form of researcher bias. Random error is meant to represent uncontrollable sources of deviations between observed and true values. Thus, terms such as "bias" used in the qualitative methodological literature would not likely approximate the same meaning as error within quantitative methods. However, given that some qualitative authors discussed the accuracy of representations in the context of trustworthiness and validity, the term "representative" might be one to explore in future work. Would a qualitative researcher take an "unrepresentative" observation to imply greater amounts of error associated with those observations? Moreover, what makes a "representative" observation in qualitative research? In my investigation of "trustworthiness" I found that some authors discussed accuracy in terms of accurate representations of an individual's experience. Would there perhaps be some amount of "noise" involved in one's *interpretations* of experience, or would that "noise" be considered as part of the experience? It appears that an application of "error" for qualitative research is not clear-cut and that will likely depend on the meaning of "truth" adopted by researchers.

In sum, my investigations of the qualitative methods literature led to the following key points regarding the role of "error" in qualitative research:

- Many qualitative researchers approach their objects of study from a perspective on truth that is fundamentally in opposition to the objectivist viewpoint tied to the concept of "error" in quantitative methods.
- Given that the "error model" stemming from statistics relies on the central notion that a true score is fixed and independent of observer influence, a direct application of the error concept to qualitative research would need to adopt this line of thinking.
- Although qualitative researchers may not accept a strong objectivist view of "true scores," they seemingly still care about the accuracy of their results in the sense that they want to accurately represent the subjective experiences that they describe.
- A variation of "error" that speaks to the misrepresentation of people's experiences might be suitable for qualitative research. This form of "error" might be described using alternative terms in the same way that "reliability" has been described in alternative terms.

These points highlight potentially fundamental differences between quantitative and qualitative methods. At the same time, they also stress several potentially important similarities. I discuss these in the next sub-section.

6.1.3. Comparing "Error" in Quantitative and Qualitative Methods

In August 2016 I attended a panel discussion at the 124th Annual Convention of the American Psychological Association. The focus of the panel was on the role of qualitative research methods in psychological research, and, in particular, the relationship between quantitative and qualitative methods. One of the speakers on the panel was the current editor-in-chief of *Qualitative Psychology,* Ruthellen Josselson, who spoke of some of the contrasting aims of quantitative and qualitative researchers. One remark in particular focused on the notion of "error." Josselson stated that whereas quantitative researchers want to *eliminate* error, qualitative researchers want to *study* it. It appeared that in Josselson's view, quantitative and qualitative researchers have starkly different perspectives on error and on the ultimate aims of research. Indeed, this same view was repeated in some of the sources I sampled in my review of qualitative methodological literature. More recently, I surveyed my upper level undergraduate psychology students in a course on psychological assessment regarding their views on

the fixed nature of the error concept in psychology. I asked whether they believed that adopting a view of fixed true scores was reasonable given the phenomena that we study in psychology. It appeared that most of the students in the course viewed psychological phenomena as being relatively fixed and objective processes. They noted that situational factors might impact psychological objects of study (especially over time) but that it was reasonable to assume that intelligence, for example, is a fixed cognitive phenomenon that should not be impacted by an observer.

Although it seems clear that there are some important differences between how quantitative and qualitative psychological researchers view their objects of study and how they believe research questions should be addressed, my experiences talking with both quantitative and qualitative researchers lead to me to believe that there are also some important similarities. I also believe that there may be some misconceptions held by people in both methodological camps regarding the "other side." In this sub-section, I summarize and explore some of the differences and similarities that emerged through my conceptual and content analyses. Based on these analyses, several overall themes related to discourse around the methodological concept of "error" have been identified. These include how researchers conceptualize and deal with (a) accuracy, (b) truth, and (c) variation. Here, I use these overarching themes as guidelines to compare my explorations of "error" within quantitative and qualitative methods. Table 4 presents a summary of this comparison.

Table 4. A Comparison of Quantitative and Qualitative Methods Discourse

Theme	Quantitative Methods	Qualitative Methods
Accuracy	• Accuracy of observations described as an important goal. • Accuracy defined as the degree to which an observation reflects a true value. • The aim of psychometrics described as being the determination of accuracy of observed values (Gulliksen, 1950). • Uncertainty is quantified.	• Accuracy of interpretations described as an important goal. • Researchers' interpretations of data should accurately reflect the lived experiences of participants. • Accuracy is related to the "trustworthiness" of qualitative research. • Trustworthiness can be accomplished through member checking, reflexivity, and transparency.
Truth	• Truth (or the "true score") is fixed and objective. • The truth is objective in the sense that it exists independently of observers. • Truth is measurable/quantifiable. • Truth can be separated from context. • Truth is singular.	• Views on truth depend on the philosophical perspective adopted by researchers. • Most commonly, truth is viewed as being multifaceted and subjective. • Truth does not exist independently from observers. • Truth is not always measurable and/or quantifiable. • Truth cannot be separated from context. • Multiple truths can exist; truth is plural.
Variation	• Variation defined in terms of dispersion (distances) around a central value (typically the mean). • Variation can be described at three different levels (intra-level in-vacuo, intra-level repeated measurements, and inter-level). • Variation often interchangeable with "error." • Variation is used to create complex mathematical models that are taken as representations of nature.	• Variation defined in terms of subjectivity (sometimes, variation is taken to be analogous to subjectivity). • Variation not defined mathematically. • Complexity of variation an important factor. • Variation not interchangeable with "error." • Different levels of analysis not described.

The pursuit of accuracy was a consistent theme in my historical analysis of the statistical concept of "error." Interestingly, I also found it to be discussed within the qualitative methodological literature. The accuracy of qualitative researchers' representations of subjective experience is taken to be an important indicator of the quality of qualitative studies. In fact, authors have discussed issues related to the accuracy of interpretations based on qualitative research methods in different ways. For example, in a paper titled, "That's Not What I Said," Borland (1991) explored variations in interpretation and meaning. She examined the notion that a researcher's interpretation

of an event might differ from a participant who experienced that event. Moreover, she acknowledged that narratives of events are not fixed and are likely to change over time. This may lead to an issue of participants feeling *misrepresented* by the interpretations that researchers make based on descriptions or observations of participant experience. Clark and Sharf (2007) further pointed out that issues of misrepresentation are, in fact, issues of ethics for qualitative researchers. Qualitative researchers often deal with very personal and difficult topics. Clark and Sharf (2007) explored issues around presenting "truth" when a participant's privacy and security might be compromised. Qualitative researchers might struggle with portraying accurate representations of personal experience when parts of that experience cannot be shared due to ethical reasons.

Thus, it appears that accuracy is a notion that is relevant for both quantitative and qualitative research. However, there are differences in terms of *why* it might be relevant and *how* it might be addressed. In statistical analyses, accuracy is modelled quantitatively, with a researcher's level of "certainty" being akin to the average amount of error associated with observations. In qualitative research, accuracy of results is related to the notion of "trustworthiness" and the focus is on accurate representation rather than on quantitative modelling of distances between observed values and true values. Moreover, the notion of "truth" varies between quantitative and qualitative research. Although qualitative research methodologists in my sample of articles pointed out that qualitative research can be conducted through a lens that ascribes "objectivist" values to objects under study, by far, "truth" is typically viewed in a subjective way in qualitative research. That is, statistical methods presume that truth is fixed, exists independent of observers, and is singular. In qualitative research, truth can be dynamic, dependent on context and interpretation, and is often conceived of as being plural. One's conception of truth will therefore has implications for what is meant by "accuracy." If one conceives of truth as being fixed, independent of observation, and numerically representable, then perhaps a statistical modelling approach is appropriate for increasing accuracy and capturing such truth. However, qualitative researchers that view truth in a much more subjective manner rely on techniques such as transparency and reflexivity in their pursuits of accuracy.

Although truth may be conceptualized differently by quantitative and qualitative researchers, Randall and Phoenix (2009) remind us that both kinds of researchers face similar challenges in terms of accuracy of representation. A research participant that

might not accurately recall an event during a qualitative interview session might just as well inappropriately rate themselves on a self-report questionnaire. A quantitative researcher might deal with this by using concepts such as systematic or random error, while a qualitative researcher might deal with this issue by considering subjectivity in interpretation. Nonetheless, it remains an issue to be dealt with in both domains. In addition, although quantitative researchers use statistical models that presume a fixed and objective true value, such researchers would not deny the subjectivity in interpretation. The difference between quantitative and qualitative research, however, lies in the fact that quantitative researchers believe that such subjectivity in interpretation can be eliminated through rigorous experimental control and absolution of observer bias; whereas, qualitative researchers *incorporate* subjectivity in interpretation as part of their methods. Importantly, quantitative researchers thus do not necessarily aim to *eliminate* error as some qualitative researchers might attest. Rather, quantitative researchers aim to eliminate observer bias and subjectivity. Random error, on the other hand, is often used as a tool for making inferences (e.g., in the method of least squares described in Chapter 1), particularly when it is interpreted as variation at the aggregate level. I return to this point later in this sub-section.

It is also important to note here that certain statistical and qualitative *methods* are built on foundations that presume either an “objectivist” or “subjectivist” view of objects under study. However, it is likely that the average *researcher*, whether quantitative or qualitative, does not hold one or the other view, but rather a mix of both (Court, 2013). That is, there can be subjectivity in objectivity and vice versa, and it would be hard to find a reasonable researcher that would deny a degree of “realism” regarding their objects of study. Thus, perhaps it is not up to the methodological tools that we use to dictate the research process, but rather for the researcher to *reflectively choose* methodological tools that are appropriate given their research questions.

One issue that should be reflected upon is the level of analysis at which a research question operates. Levels of analysis appear to be more explicitly considered in quantitative and statistical methodological tools, although they are not irrelevant to qualitative research. I identified three levels of analysis at which quantitative researchers potentially conceptualize variation. The first, is the intra-individual level in which repeated observations from the same person are conceptualized “in vacuo” or independent of time. The second, is the intra-individual level in which repeated observations from the

same person are conceptualized over time. And, the third is the inter-individual level in which observations are taken from different people. How “error” is defined will vary at these three different levels. Psychometric theory typically starts at the “in vacuo” intra-individual level and theorizes that the average value of this propensity distribution is equivalent to a person’s true score. Error is then defined as a distance between an observation and a person’s true score. However, when moving to the other 2 levels of analysis, distinctions between “error” and “variation” become tricky. For example, depending on the statistical tool being used, distances from the mean at the inter-individual level can be viewed as “between-person error,” or as “individual differences” (i.e., variation), or both.

Importantly, regardless of interpretation, distances between a mean value and some observed score are almost always used to mathematically “model” affairs in nature. This is important to note because quantitative researchers do not, in fact, aim to *eliminate* error; rather, they aim to *use* it, typically to make inferences about average values to broader populations of scores. Some statistical procedures do so through processes in which error is minimized; however, if all error were to be eliminated there would arguably be no need for inferential statistics or psychometric theory. Previously, I described how some qualitative researchers held the impression that quantitative researchers aim to eliminate error, whereas qualitative researchers aim to study it. In qualitative research, variation is described in terms of subjectivity and is not mathematically modelled. However, based on my analyses, I argue that these qualitative researchers are in fact misunderstanding the concept of “error” in statistics. Error terms are not replacements for subjectivity. Rather, they are meant to capture misrepresentations of phenomena that come about through random chance. Any quantitative researcher using statistical methods to study psychological phenomena would be interested in factors that explain individual differences and/or subjectivity related the objects under study. Moreover, quantitative researchers do not advocate modelling such meaningful differences as random error. Rather, such differences might be treated as extraneous factors that should be accounted for in the design of a research study. These extraneous factors might indeed introduce *bias* into study results, but they are not taken to be akin to random error.

Thus, it appears that what some qualitative researchers might mean when they reference error in quantitative methods is what actually would be considered to be *bias*

by quantitative researchers. For example, in a previous chapter, I described how Norris (1997) conflated the terms “error” and “bias.” Based on my analyses, I argue that both quantitative and qualitative researchers are interested in minimizing error in that they both aim for accurate representations. I also argue that both kinds of researchers are interested in individual differences and subjectivity. However, how these various concepts are treated and conceptualized from ontological and epistemological standpoints will vary from discipline to discipline and from researcher to researcher. That is, an “accurate representation” for a quantitative researcher might mean obtaining the value of a fixed true score; whereas, an “accurate representation” for a qualitative researcher might mean portraying the subjective experiences of individuals in an ethical way. Nonetheless, quantitative and qualitative researchers appear to have many commonalities. Both rely on variation in objects of study to conduct research and make inferences (Wertz, 1986), both aim to represent truth (however that “truth” might be conceptualized) in meaningful ways, both aim to conduct quality research in terms of consistency, validity, and trustworthiness, and both engage in the pursuit of accuracy.

6.2. Implications and Conclusions

I began my project with three overarching research aims. I questioned the role and meaning of error in quantitative psychological research and wondered whether error had any relevance for qualitative psychological research. I was also curious about comparisons between these two supposedly opposing domains of inquiry. In exploring each of these aims, I also wanted to explicate the potential implications of my analyses for methods-use in psychology. Thus, in this final section I discuss some of the implications that arise from my overall project, and, in particular, from my comparisons of quantitative and qualitative research methods. I organize these implications into 3 categories: (a) the treatment of variation; (b) the role of philosophical commitments; and (c) education.

6.2.1. The Treatment of Variation

In quantitative research, how one understands “error” has implications for how variation (distances between observed scores and average scores) might be treated and interpreted. However, it appears that it is not always clear what level of analysis

psychologists aim to make inferences from. One implication of my current analyses speaks to the importance of considering the kinds of inferences and interpretations that researchers want to make based on their findings. As I have shown, “error” and “variation” will have different meanings depending on level of analysis. Researchers should consider whether they aim to make inferences that are true on a general level (i.e., common to all), or if they aim to make inferences that are true on average (Bakan, 1967). For example, if, like Wundt, researchers aim to explicate what is true of individual experience, then they should consider working with variation at the intra-individual level. However, if, like Galton, researchers aim to make inferences that are true on average about a given population, then they may work at the inter-individual level.

Throughout the 20th century, it has been unclear whether psychological researchers have adequately addressed issues concerning level of analysis. These concerns have been raised by several scholars including Lamiell (2003), Gigerenzer (1987), and Molenaar (2004). These authors have argued that psychological researchers continuously rely primarily on statistical tools that model inter-individual variation to answer questions better suited to studies of intra-individual variation. More recently, psychological researchers that have shown explicit interest in capturing intra-individual changes over time have argued for uses of statistical methods that allow for the modelling of both intra- and inter- individual variation. For example, in developmental psychology, theories of dynamic and embodied development that call on researchers to study phenomena from multiple levels of analysis have pushed forth the adoption of methods that presumably allow for inferences at multiple levels (Diehl, Hooker, & Sliwinski, 2015). These methods fall under the umbrella of “multilevel modelling” in which repeated measurements from the same person obtained over time are conceptualized as being nested within the individual from whom the measurements were taken. This individual is then conceptualized as being nested within a larger group of individuals taken to represent the population of interest. Thus, it appears that these methods allow for the “modelling” of both intra- and inter- levels of variation. However, as Molenaar (2015) has pointed out, the use of such models to study intra-level variability presents yet another example of psychologists’ misunderstanding of levels of analysis. In multilevel modelling, inferences are made based on differences across people; inferences are rarely made at the individual level. That is, such modeling of repeated measures is always conducted through the pooling of data values across the individual

participants who have been repeatedly observed rather than through the pooling of the repeated observations across time (Molenaar, 2015).

Given the above, what methods should a psychologist who wants to make claims at the intra-individual level use? One might draw inspiration from *N*=1 studies in which the same person is repeatedly observed over time and inferences are made based on the pooling of these repeated observations (Diehl et al., 2015). This method would certainly be relevant for researchers interested in intra-individual variation. Here, "error" would be considered as a distance between a repeated observation and a person's true score. In addition, Grice (2015) developed observation oriented modeling as a method that focuses on the detections of patterns rather than on aggregate level statistics such as means and variances. However, researchers may also find that if they are open to experimenting with a broader variety of research techniques they may be able to better address their research questions. More specifically, a combination of statistical analyses such as multilevel modeling and qualitative analyses that allow for a more in-depth analysis of individual experience might be appropriate. Based on my historical analysis of the concept of "error," I suspect that one of the reasons that psychologists continuously use analyses based on inter-level variation to answer research questions about intra-level variation is because of the lack of statistical tools that can adequately handle the study of individual experience. Greater openness to qualitative methods may help to fill this gap in psychological research.

6.2.2. Philosophical Commitments

A second implication of the current work is the importance of considering the philosophical foundations that underlie methodological practices. It is clear from my historical analysis that the statistical concept of error stems from a model that presumes an objectivist view of objects under study. The notion of a "fixed" true score independent of observer influence speaks to this objectivist foundation. At the same time, it is clear that the domain of qualitative research is closely tied to philosophical outlooks that favour a constructivist/interpretivist perspective in which the acknowledgement of subjectivity is of central interest to the research process. Indeed, psychological researchers should be fully aware of the philosophical foundations of methodological tools and should carefully consider their implications for research studies. However, I am

hesitant to recommend that researchers only utilize tools that fit their own philosophical viewpoints because it may lead to a form of dogmatism in research practice.

Such limitation would only lead to greater misunderstanding of methods and less opportunity for interdisciplinary collaboration. This is a view shared by many methodologists who advocate against dogmatism in method-use and support the "mixing" of quantitative and qualitative research techniques. Authors such as Howe (1988), Biesta (2010), and Onwuegbuzie and Leech (2005) have argued that the perceived incompatibility of quantitative and qualitative research methods is not justified. Both domains of research encompass a vast array of methodological techniques, and, as I have shown through my historical and content analyses, both domains of research are concerned with accuracy in terms of representing truth in an appropriate manner. Acknowledging that methodological tools were developed based on a particular philosophical foundation does not justify the limitation of use of a tool to individuals who hold the same philosophical beliefs. For example, a researcher who holds the view that their objects of study exist independent of observer influence may find variations of thematic or narrative analysis useful in answering their research questions. In addition, a researcher who adheres to a phenomenological perspective of knowledge might conduct an analysis of variance to inform their understanding of observations obtained from a large sample of people. In both quantitative and qualitative research, methodological techniques do not speak for themselves. They always require a researcher to make judgements about the findings. Moreover, my analyses of both quantitative and qualitative methodology have shown that there are shared aims between these two seemingly opposing domains. Thus, researchers should not limit themselves to methodological tools that comply with specific philosophical commitments.

It should additionally be noted that several lines of exploration related to the philosophical foundations of research methods and psychology have been touched upon in the current project without extended exploration. In an attempt to remain focused on my initial aims, I have actively avoided delving deeper into these issues. Such issues include topics such as, similarities and differences between psychological and physical phenomena, issues pertaining to what constitutes an "unobservable" phenomenon and whether psychological phenomena can usefully be considered as "unobservable," and the distinction between objectivism and subjectivism in science. These issues represent

important areas of inquiry in relation to the topics examined in the current work and should be considered more fully in future work.

6.2.3. Education

Finally, in order to promote better understanding of psychological research methods, education of psychological researchers must be taken into consideration. Statistical textbooks geared at teaching methods in psychology rarely emphasize the meanings of concepts or the underlying logic of statistical tools. Conceptual understanding takes a backseat to applied and operational knowledge. Students spend weeks learning programming languages and statistical software and yet exit courses without proper understanding of statistical concepts. Philosophical assumptions inherent in the definitions of statistical concepts are not made explicit. For example, a popular intermediate level statistics text by Howell (2010) defines a "variable" as "a property of an object or event that can take on different values," providing "self-confidence" as an example (p. 4). Nowhere is it acknowledged that this definition and example of "variable" implies a philosophical position in which a variable is taken to be *equivalent* to a psychological object. Moreover, textbooks of psychological assessment and measurement fail to explicate the point that true score theory is theorized at the person-level (e.g., Kaplan & Saccuzzo, 2018). Rather, emphasis is placed on defining different "types" of reliability and validity. As I have shown through my analysis of the "error" concept, misunderstandings of methodological concepts and theories can have implications for improper use of statistical tools. That is, if students do not fully understand statistical and qualitative research concepts and their underlying philosophical and theoretical assumptions they may blindly misuse such concepts. For example, a misunderstanding of "error" and how the concept functions at different levels of analysis has mistakenly led developmental researchers to use techniques such as multi-level modeling to answer questions about intra-level variation (Molenaar, 2015).

Moreover, if students are not exposed to a diverse range of methods in their undergraduate and graduate studies they will likely continue to use the same two or three methodological tools that they are most familiar with to answer every research question. What's more concerning, they might begin to form their research questions to suit methods rather than to choose methods that suit research questions. Alternatively, they might use certain methods solely because they are the conventional tools of their

field. For example, in my analysis of qualitative methodological papers, I observed that many qualitative researchers felt compelled to examine reliability solely because it was a convention of their research domain. This therefore also points to the need for wider acceptance of methodological diversity at the systemic level. Courses in psychological research methods should place greater emphasis on conceptual understanding in psychology methods courses, as well as greater emphasis on diversity in methods-use.

6.2.4. Concluding Remarks

I began the current project with an interest in clarifying the meaning of "error" in uses of statistical methods within the field of psychology. This interest stemmed from my questioning of measurement practices in psychology and whether statistical analytic tools are most appropriate for answering psychological research questions. I was drawn to the concept of "error" because of what it presumably represented in my naïve view of the term. That is, I viewed it as representing inaccuracies associated with psychological measurement. My initial view of "error" is indeed related to the meaning(s) of the concept for psychology; however, it does not tell the entire story of how central the "error" concept is for statistics in psychology. Although it is not always clear what researchers mean when they invoke the notion of "error," it is clear that it is the foundation of the most popular statistical tools of psychology (e.g., psychometrics, ANOVA, regression analysis).

My exploration of "error" and my questioning of measurement in psychology led me to qualitative methods. Qualitative research was proposed to me as an "alternative" to measurement in psychology. This made me curious about whether the concept of "error" played a role in qualitative methods. My exploration of qualitative methodological papers and their discussions of "reliability" and "trustworthiness" led me to discover that qualitative and quantitative researchers, although divided in their views on truth, share some common aims. Both domains of research are concerned with the quality of their studies in terms of "accurately" representing psychological and social phenomena. This shared aim led me to conclude that "error" is relevant for both areas of practice insofar as it represents deviations from truth (regardless of how "truth" might be defined). Thus, both quantitative and qualitative researchers struggle with issues of accuracy and truth representation. Both also require the development of tools to establish the consistency and trustworthiness of research results.

Given the above considerations, I suggest that psychological researchers begin to emphasize conceptual understanding of statistical and methodological research tools. This would mean that researchers understand the philosophical assumptions that underlie the foundations of central statistical concepts such as “error.” In addition, researchers should understand that methodological tools are tied to specific histories. These histories are not merely linear narratives of tool advancement; rather, they help to explain the conceptual significance of such tools, and therefore, have important implications for their uses. Finally, researchers should not pick and choose their methodological tools based on pre-conceived assumptions about knowledge and/or reality as this can potentially lead to dogmatism in methods-use. Rather, an openness to diverse research methods (both quantitative and qualitative) should lead to better methods understanding and more appropriate methods-use.

References

Abelson, A. R. (1911). The measurement of mental ability of “backward children.” *British Journal of Psychology, 4,* 268-314.

Allport, G. W. (1942). *The use of personal documents in psychological science* (prepared for the Committee on the Appraisal of Research; Bulletin #49). New York: Social Science Research Council.

Bakan, D. (1967). *On method: Toward a reconstruction of psychological investigation.* San Francisco, CA: Jossey-Bass.

Batten, A. H. (2015). A brief history of error. *Journal of Astronomical History and Heritage, 18*(2), 116-122.

Berka, K. (1983*). Measurement: Its concepts, theories, and problems*. Boston: Reidel.

Biesta, G. (2010). Pragmatism and the philosophical foundation of mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods research for the social and behavioral sciences* (2nd ed., pp. 95-118). Thousand Oaks, CA: Sage Publications.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s ability. In F. Lord & M. Novick (Eds.)*, Statistical theories of mental test scores*. (pp. 397-479). Reading: Addison-Wesley.

Birnbaum, A. (1969). Statistical theory for logistic mental test models: A prior distribution ability. *Journal of Mathematical Psychology*, 6, 258-276.

Boring, E. G. (1929). *A history of experimental psychology*. New York, NY: The Century Co.

Borland, K. (1991). “That’s not what I said:” Interpretive conflict in oral narrative research. In S. B. Gluck & D. Potai (Eds.), *Women’s words: The feminist practice of oral history.* New York: Routledge.

Braun, V. & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research In Psychology, 3,*(2), 77-101.

Brinkmann, S., Jacobsen, M. H., & Kristiansen, S. (2014). Historical overview of qualitative research in the social sciences. In P. Leavy (Ed.), *The Oxford handbook of qualitative research.* New York: Oxford University Press.

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

Burt, C. (1955). Test reliability estimated by analysis of variance. *British Journal of Statistical Psychology, 8* 103-118.

Clark, M. C. & Sharf, B. F. (2007). The dark side of truth(s): Ethical dilemmas in researching the personal. *Qualitative Inquiry, 13*(3), 399-416.

Comte, A. (1975). *Auguste Comte and positivism: The essential writings* (G. Lenzer, Ed.). New York, NY: Harper Torchbooks.

Condon, D. M., Wilt, J., & Revelle, W. (2011). Individual differences and differentiall psychology: A brief history and prospect. In T. Chamorro-Premuzic, S. von Stumm, & A. Furnham (Eds.), *The Wiley-Blackwell handbook of individual differences.* (pp. 1-38). Blackwell Publishing Ltd. doi: 10.1002/9781444343120

Cornfield, J., & Tukey, J. W. (1956). Average values of mean squares in factorials. *Annals of Mathematical Statistics, 27,* 907-949.

Court, D. (2013). What is truth in qualitative research? Why is this important for education? *Educational Practice and Theory, 35*(2), 5-14.

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

Daston, L. & Galison, P. (2007). *Objectivity*. Brooklyn, NY: Zone Books.

Danziger, K. (1987). Statistical method and the historical development of research practice in american psychology. In L. Kruger, L. G. Gigerenzer, & M. S. Morgan (Eds.), *The probabilistic revolution: Volume 2: Ideas in the sciences* (pp. 35-48). Cambridge, MA: MIT Press.

Denis, D. J. (2001). The origins of correlation and regression: Francis Galton or Auguste Bravais and the error theorists? *History and Philosophy of Psychology Bulletin, 13*(2), 36-44.

Dewey, J. (1910). *How we think.* Amherst, NY: Promtheus Books.

Diehl, M., Hooker, K., & Sliwinski, M. J. (2015). A brief historical overview of intraindividual variability research across the life span. In M. Diehl, K. Hookher, & M. J. Sliwinski (Eds.), *Handbook of intraindividual variability across the life span.* (pp. 3-10). New York: Routledge.

Dilthey, W. (2002). *The formation of the historical world in the human sciences* (Willhelm Dilthey Selected Works Vol. 3). R. A. Makkreel & F. Rodi (Eds.). Princeton: Princeton University Press.

Dreyfus, H. & Taylor, C. (2015). *Retrieving realism.* Cambridge, MA: Harvard University Press.

Fechner, G. T. (1860). *Elemente der psychophysik*. Leipzig, Germany: Breitkopf & Hartel. English translation by H. E. Adler (1966), *Elements of psychophysics.* (D. H. Howes & E. G. Boring, Eds.) New York, NY: Holt, Rinehart & Winston.

Fechner, G. T. (1966). *Elemente der psychophysik*. Leipzig, Germany: Breitkopf & Hartel. English translation by H. E. Adler, *Elements of psychophysics.* D. H. Howes & E. G. Boring, Eds.) New York, NY: Holt, Rinehart & Winston. (Original work published 1860).

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Philosophical transactions of the Royal Society of Edinburgh, 52*, 399-433.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical transactions of the royal society, 222*, 309-368.

Fisher, R. A. (1925). *Statistical methods for research workers.* Edinburgh: Oliver and Boyd.

Flicker, S. (2004). “Ask me no secrets, I’ll tell you no lies”: What happens when a respondent’s story makes no sense. *The Qualitative Report, 9*(3), 528-537.

Freud, S. (1965). *The interpretation of dreams.* New York: Basic Books. (Original work published 1900).

Gadamer, H. G. (1960). *Wahrheit und method*. Tubingen: Breitkopf & Hartel. English translation by J. Weinsheimer & D. G. Marshall (1975), *Truth and Method* (2nd ed.). London: Continuum.

Galton, F. (1869). *Hereditary genius*. London, UK: Macmillian.

Galton, F. (1875). Statistics by intercomparison, with remarks on the law of frequency of error. *Philosophical Magazine 4th series 49*(322), 33-46.

Galton, F. (1888). Co-relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society, 45,* 135-145.

Gauss, C. F. (1809). *Theoria motus corporum celestium.* Hamburg: Pertheset Besser. Translation by C. H. Davis (1857), *Theory of motion of the heavenly bodies moving about the sun in conic sections*. Boston: Little, Brown.

Gauss, C. F. (1857). *Theory of motion of the heavenly bodies moving about the sun in conic sections.* (C. H. Davis, Trans.). Boston, MA: Little, Brown. (Original work published 1809)

Gergen, K. J. (2018). Qualitative psychology and the new pluralism. In B. Schiff (Ed.). *Situating qualitative methods in psychological science.* New York: Routledge.

Gergen, K. J., Josselson, R., & Freeman, M. (2015). The promises of qualitative inquiry. *American Psychologist, 70*, 1-9. doi: 10.1037/a0038597

Gigerenzer, G. (1987). Probabilistic thinking and the fight against subjectivity. In L. Kruger, L. G. Gigerenzer, & M. S. Morgan (Eds.), *The probabilistic revolution: Volume 2: Ideas in the sciences* (pp. 11-34). Cambridge, MA: MIT Press.

Gigerenzer, G., Swijtink, Z. Porter, T. Daston, L., Beatty, J., & Krüger, L. (1989). *The empire of chance: How probability changed science and everyday life*. Cambridge: Cambridge University Press.

Grice, J. W. (2015). From means and variances to persons and patterns. *Frontiers in Psychology, 24.* doi: 0.3389/fpsyg.2015.01007.

Guba, E. G. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. *Educational Communication and Technology, 29*(2), 75-91.

Guba, E. G. & Lincoln, Y. S. (1981). *Effective evaluation.* San Francisco: Jossey-Bass.

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley.

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

Guttman, L. (1950). Relation of scalogram analysis to other techniques. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction.* (pp. 172-212). Princeton: Princeton University Press.

Guttman, L. (1953). A special review of Harold Gulliksen, “Theory of mental tests.” *Psychometrika, 18*, 123-130.

Hald, A. (2007). *A history of parametric statistical inference from Bernoulli to Fisher.* New York: Springer.

Heidegger, M. (1927). *Sein und zeit*. De Gruyter. English translation by J. Stambaugh (1996), *Being and time.* New York: State University of New York Press.

Heidelberger, M. (1987). Fechner’s indeterminism: From freedom to laws of chance. In L. Kruger, L. J. Daston, & M. Heidelberger (Eds.), *The probabilistic revolution: Volume 1: Ideas in history* (pp. 117-156). Cambridge, MA: MIT Press.

Howe, K. R. (1988). Against the quantitative-qualitative incompatibility thesis or dogmas die hard. *Educational Researcher, 17*(8), 10-16.

Howell, D. C. (2010). *Statistical methods for psychology* (7th ed.). Belmont, CA: Cengage Wadsworth.

Hsieh, H. & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative health research*, 15(9), 1277-1288.

Husserl, E. (1954). *Die Krisis der europaischen Wissenschaften und die tranzendentale Phanomenologie.* Haag: Martinus Nijhoff.

Jahoda, G. (2015). Quetelet and the emergence of the behavioral sciences. *SpringerPlus, 4*(473), 1-10.

James, W. (1902). *The varieties of religious experience: A study in human nature.* Auckland: The Floating Press.

James, W. (1907). *Pragmatism: A new name for some old ways of thinking.* Cambridge, MA: Harvard University Press.

Kaplan, R. M. & Saccuzzo, D. P. (2018). *Psychological testing: Principles, applications, and issues* (9th ed.). Boston, MA: Cengage Learning.

Kelley, T. L. (1921). The reliability of test scores. *Journal of Educational Research, 3*(5), 370-379.

Kelley, T. L. (1924). Note on the reliability of a test: A reply to Dr. Crum's criticism. *The Journal of Educational Psychology, 14*(4), 193-204.

Kelley, T. L. (1927). *Interpretation of educational measurements.* New York: World Book.

Kohlberg, L. (1981). *Essays on moral development* (Vol. 1). San Francisco: Harper & Row.

Krantz, D. H. (1991). From indices to mappings: The representational approach to measurement. In D. Brown & J. Smith (Eds), *Frontiers of mathematical psychology: Essays in honor of Clyde Coombs* (pp. 1-52). New Yorkw: Springer-Verlag.

Kruger, L. (1987). The slow rise of probabilism: Philosophical arguments in the nineteenth century. In L. Kruger, L. J. Daston, & M. Heidelberger (Eds.), *The probabilistic revolution: Volume 1: Ideas in history* (pp. 59-90). Cambridge, MA: MIT Press.

Kruger, L., Daston, L. J., & Heidelberger, M. (Eds.). (1987). *The probabilistic revolution: Volume 1: Ideas in history.* Cambridge, MA: MIT Press.

Kruger, L., Gigerenzer, G., & Morgan, M. S. (Eds.). (1987). *The probabilistic revolution: Volume 2: Ideas in the sciences.* Cambridge, MA: MIT Press.

Kuder, G. F. & Richardson, M.W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151-160.

Kuhn, T. H. (1998). The natural and the human sciences. In E. D. Klemke, R. Hollinger, & D. W. Rudge (Eds.), *Introductory readings in the philosophy of science* (pp 128-134). Amherst, NY: Prometheus.

Lamiell, J. T. (2003). *Beyond individual and group differences: Human individuality, scientific psychology, and William Stern's critical personalism.* Thousand Oaks, CA: Sage.

Lamiell, J. T. (2013). Statisticism in personality psychologists' use of trait construcgts: What is it? How was it constructed? Is there a cure? *New Ideas in Psychology, 31,* 65-71.

Lamiell, J. T. (2018). Some historical perspectives on the marginalization of qualitative methods within mainstream scientific psychology. In B. Schiff (Ed.). *Situating qualitative methods in psychological science.* New York: Routledge.

Latour, B. (2000). When things strike back – a possible contribution of "science studies" to the social sciences. *British Journal of Sociology, 51*, 107-123.

Lawley, D. N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburg, 61*, 273-287.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction.* (pp. 362-412). Princeton: Princeton University Press.

Levitt, H. M., Bamberg, M., Creswell, J. W., Frost, D. M., Josselson, R., & Suarez-Orozco, C. (2018). Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: The APA publications and communications board task force report. *American Psychologist, 73* (1), 26-46.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635-694.

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

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

Madill, A., Jordan, A., & Shirley, C. (2000). Objectivity and reliability in qualitative analysis: Realist, contextualist and radical constructionist epistemologies. *British Journal of Psychology, 91*, 1-20.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. New York: Routledge.

Mellenbergh, G. J. (2011). *A conceptual introduction to psychometrics: Development, analysis, and application of psychological and educational tests.* The Netherlands: Eleven International.

Merlaeu-Ponty, M. (1945). *Phenomenology of perception*. London: Routledge.

Michell, J. (2003). The quantitative imperative: Positivism, naïve realism and the place of qualitative methods in psychology. *Theory & Psychology, 13,* 5-31.

Michell, J. (2004). *Measurement in psychology: A critical history of a methodological concept*. New York, NY: Cambridge University Press.

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement, 2*(4), 201-218.

Molenaar, P. C. M. (2015). The future of analysis of intraindividual variation. In M. Diehl, K. Hookher, & M. J. Sliwinski (Eds.), *Handbook of intraindividual variability across the life span.* (pp. 343-357). New York: Routledge.

Norris, N. (1997). Error, bias and validity in qualitative research. *Educational Action Research, 5*(1), 172-176.

Nunnally, J. M. (1970). *Introduction to psychological measurement.* New York, NY: McGraw-Hill.

NVivo Qualitative Data Analysis Software (2017). Version 11. QSR International Pty Ltd.

Onwuegbuzie, A. J., & Leech, N. L. (2005). On becoming a pragmatic researcher: The importance of combining quantitative and qualitative research methodologies. *International Journal of Social Research Methodology, 8,* 375-387.

Oxford University Press (2018). Reliability. Retrieved from https://en.oxforddictionaries.com/definition/reliability

Pearson, K. (1892). *The grammar of science.* London: Walter Scott.

Pearson, K. (1896). Mathematical contributions to the theory of evolution – III. Regression, heredity and panmixia. *Philosophical Transactions, A, 187*, 252-318.

Pearson, K. (1904). On the laws of inheritance in man. II. On the inheritance of the mental and moral characters in man, and its comparison with the inheritance of physical characters. *Biometrika, 3,* 131-190.

Piaget, J. (1932). *The moral judgment of the child.* Glencoe, IL: Free Press.

Porter, T. M. (1986). *The rise of statistical thinking: 1820-1900.* Princeton, NJ: Princeton University Press.

Quetelet, A. (1842). *A treatise on man and the development of his faculties.* Edinburgh: Chambers. (Original work published 1835)

Randall, W. L. & Phoenix, C. (2009). The problem with truth in qualitative interviews: Reflections from a narrative perspective. *Qualitative Research in Sport and Exercise, 1*(2), 125-140.

Read, C. B. (1985). Normal distribution. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 6, pp. 347-359). Toronto, ON: Wiley.

Robinson, D. N. (2001). Sigmund Koch – philosophically speaking. *American Psychologist, 56,* 420-424.

Sale, J. E. M., Lohfeld, L. H., & Brazil, K. (2002). Revisiting the quantitative-qualitative debate: Implications for mixed-methods research. *Quality & Quantity, 36,* 43-53. doi: 10.1023/A:1014301607592

Shenton, A. K. (2004). Strategieis for ensuring trustworthiness in qualitative research projects. *Education for Information, 22*, 63-75.

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

Simpson, T. (1755). A letter to the Right Honorable George Earl of Macclesfield, President of the Royal Society, on the advantage of taking the mean of a number of observations, in practical astronomy. *Philosophical Transactions of the Royal Society of London, 49,* 82-93.

Silverman, D. (1993). *Interpreting qualitative data: Methods for analyzing talk, text, and interaction.* London: Sage.

Slaney, K. L. (2001). On empirical realism and the defining of theoretical terms. *Journal of Theoretical and Philosophical Psychology, 21*(2), 132-152.

Slaney, K. L. (2006). *The logic of test analysis: An evaluation of test theory and a proposed logic for test analysis* (Unpublished doctoral dissertation). Simon Fraser University, Burnaby BC.

Spearman, C. (1904a). The proof and measurement of association between two things. *American Journal of Psychology, 15, 72-101.* doi: 10.2307/1422689

Spearman, C. (1904b). "General intelligence," objectively determined and measured. *American Journal of Psychology, 15,* 201-293. doi:10.2307/1412107

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology, 18*, 160-169.

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

Stanley, J. C. (1962). Analysis-of-variance principles applied to the grading of essay tests. *Journal of Experimental Education, 30,* 279-283.

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). (pp. 356-442). Washington, D.C.: American Council on Education.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677-680.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: The Belknap Press of Harvard University Press.

Sutcliffe, J. P. (1965). A probability model for errors of classification. I. General conditions. *Psychometrika, 30,* 73-96.

Thorndike, R. L. (1964). Reliability. In *Proceedings of the 1963 Invitational Conference on Testing Problems* (pp. 23-32). Princeton, NJ: Educational Testing Service.

Thorndike, R. L. (1966). Reliability. In A. Anastasi (Ed.). *Testing problems in perspective.* Washington, D.C.: American Council on Education.

Thurstone, L. L. (1932). *The reliability and validity of tests.* An Arbor, MI: N. p.

Traub, R. E. (1994). *Reliablity for the social sciences: Theory and applications.* Thousand Oaks, CA: Sage Publications.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice, 16*(4), 8-14. doi: 10.1111/j.1745-3992.1997.tb00603.x

Walker, H. M. (1929). *Studies in the history of statistical method*. Balitmore: Williams & Wilkins.

Watt, D. (2007). On becoming a qualitative researcher: The value of reflexivity. *The Qualitative Report, 12*(1), 82-101.

Wertz, F. J. (1986). The question of the reliability of psychological research. *Journal of Phenomenological Psychology, 17*(2), 181-205.

Wertz, F. J., Charmaz, K., McMullen, L M., Josselson, R. , Anderson, R., & McSpadden, E. (2011). *Five ways of doing qualitative analysis: Phenomenological psychology, grounded theory, discourse analysis, narrative research, and intuitive inquiry.* New York: Guildford Press.

Woodworth, R. S. (1929). *Psychology.* Michigan, H. Holt & Company.

Wundt, W. (1913). *Die psychologie in kamp funs dasein [Psychology's struggle for existence]* (2nd ed.). Leipzig, Germany: Kröner.

Appendix A.

Sources Included in Qualitative Methods Literature Content Analysis: "Reliability"

Allen, M. (2017). *The SAGE encyclopedia of communication research methods.* Thousand Oaks, California: SAGE Publications.

Altheide, D. L. (2013). *Qualitative media analysis.* London, UK: SAGE Publications.

Barbour, R. S. & Morgan, D. L. (2017). *A new era in focus group research challenges, innovation and practice.* United Kingdon: Palgrave Macmillan.

Bray, M., Adamson, B., & Mason, M. (2014). *Comparative education research: Approaches and methods.* (2nd ed.). New York: Springer.

Cain, D. J., Keenan, K., & Rubin, S. (2016). *Humanistic psychotherapies: Handbook of research and practice.* (2nd ed). Washington, DC: American Psychological Association.

Campbell, J. L., Quincy, C., Osserman, J., & Pedersen, O. K. (2013). Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research, 42*(3), 294-320.

Cook, K. E. (2012). Reliability assessments in qualitative health promotion research. *Health Promotion International, 27*(1), 90-101.

Danchev, D. & Ross, A. (2014). *Research ethics for counsellors, nurses and social workers.* London, UK: SAGE Publications.

Denzin, N. K. (2015). Interpretive methods: Micromethods. In J. D. Wright (Ed.), *International encyclopedia of the social and behavioral sciences* (Vol. 12; 2nd ed.; pp. 648-651). Amsterdam: Elsevier.

Fetterman, D. M. (2015). Ethnography in applied social research. In J. D. Wright (Ed.), *International encyclopedia of the social and behavioral sciences* (Vol. 8; 2nd ed.; pp. 184-191). Amsterdam: Elsevier.

George, P. & Syrja-McNally, D. (2015). Social enquiry and action research for social work. In J. D. Wright (Ed.), *International encyclopedia of the social and behavioral sciences* (Vol. 22; 2nd ed.; pp. 269-274). Amsterdam: Elsevier.

Gubrium, J. F., Holstein, J. A., Marvasti, A., & Mckinney, K. D. (2012). *The SAGE handbook of interview research: The complexity of the craft.* Thousand Oaks, California: SAGE Publications.

Guest, G., Namey, E. E., & Mitchell, M. L. (2013). *Collecting qualitative data: A field manual for applied research.* London, UK: SAGE Publications.

King, K. A., Lai, Y., & May, S. (2017). *Research methods in language and education.* (3rd ed.). Switzerland: Springer.

Lee, Y. (2014). Insight for writing a qualitative research paper. *Family and Consumer Sciences Research Journal, 43*(1), 94-97.

MacPhail, C., Khoza, N., Abler, L., & Ranganthan, M. (2016). Process guidelines for establishing intercoder reliability in qualitative studies. *Qualitative Research, 16*(2), 198-212.

Maxwell, J.A. & Reybold, L. E. (2015). Qualitative research. In J. D. Wright (Ed.), *International encyclopedia of the social and behavioral sciences* (Vol. 19; 2nd ed.; pp. 685-689). Amsterdam: Elsevier.

Morley, C. (2015). Critical reflexivity and social work practice. In J. D. Wright (Ed.), *International encyclopedia of the social and behavioral sciences* (Vol. 5; 2nd ed.; pp. 281-286). Amsterdam: Elsevier.

Morse, J. (2015). Critical analysis of strategies for determining rigor in qualitative inquiry. *Qualitative Health Research, 25*(9), 1212-1222.

Sanjek, R. (2015). Field observational research in anthropology and sociology. In J. D. Wright (Ed.), *International encyclopedia of the social and behavioral sciences* (Vol. 9; 2nd ed.; pp. 135-139). Amsterdam: Elsevier.

Shaw, I. (2014). *Doing qualitative research in social work*. Los Angeles, CA: SAGE Publications.

Syed, M. & Nelson, S. C. (2015). Guidelines for establishing reliability when coding narrative data. *Emerging Adulthood, 3*(6), 375-387.

Udo, K. (2014). *Qualitative text analysis: A guide to methods, practices, & using software.* London, UK: SAGE Publications.

Urquhart, C. (2013). *Grounded theory for qualitative research: A practical guide.* London, UK: SAGE Publications.

Willig, C. & Stainton-Rogers, W. (2017). *The SAGE handbook of qualitative research in psychology.* London, England: SAGE Publications.

Yin, R. K. (2015). Case Studies. In J. D. Wright (Ed.), *International encyclopedia of the social and behavioral sciences* (Vol. 3; 2nd ed.; pp. 194-201). Amsterdam: Elsevier.

Appendix B.

Sources Included in Qualitative Methods Literature Content Analysis: "Trustworthiness"

Bailey, C. (2007). *A guide to qualitative field research*. Thousand Oaks, California: SAGE Publications.

Berger, R. (2015). Now I see it, now I don't: Researcher's position and reflexivity in qualitative research. *Qualitative Research, 15*(2), 219-234.

Birth, L., Scott, S., Cavers, D., Campbell, C., & Walter, F. (2016). Member checking: A tool to enhance trustworthiness or merely a nod to validation? *Qualitative Health Research, 26*(13), 1802-1811.

du Plessis, C. (2017). The method of psychobiography: Presenting a step-wise approach. *Qualitative Research in Psychology, 14*(2), 216-237.

Kornbluh, M. (2015). Combatting challenges to establishing trustworthiness in qualitative research. *Qualitative Research in Psychology, 12*, 397-414.

Le Roux, C. S. (2016). Exploring rigour in autoethnographic research. *International Journal of Social Research Methodology, 20*(2), 195-207.

Lee, Y. (2014). Insight for writing a qualitative research paper. *Family and Consumer Science Research Journal, 43*(1), 94-97.

Levitt, H. M. (2016). Qualitative methods. In J. C. Norcross, G. R. VandenBos, & D. K. Freedheim (Eds.), *APA handbook of clinical psychology: Theory and Research* (Vol. 2). Washington, DC: American Psychological Association.

McAteer, M. (2013). *Action research in education.* London: SAGE Publications.

Newman, I., Lim, J., & Pineda, F. (2013). Content validity using a mixed methods approach: Its application and development through the use of a table of specifications methodology. *Journal of Mixed Methods Research, 7*(3), 243-260.

Rodham, K., Fox, F., & Doran, N. (2015). Exploring analytical trustworthiness and the process of reaching consensus in interpretative phenomenological analysis: Lost in transcription. *International Journal of Social Research Methodology, 18*(1), 59-71.

Uwe, F. (2007). *Managing quality in qualitative research.* London, England: SAGE Publications.