STATISTICS IN AMERICAN PSYCHOLOGY: THE SOCIAL CONSTRUCTION OF EXPERIMENTAL AND CORRELATIONAL PSYCHOLOGY, 1900-1930.

by

Dale Arthur Stout

Doctor of Philosophy
University of Edinburgh
1987
For my parents and my children:

Sid & Phyllis

Aaron & Scott
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>i</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>ii</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>CHAPTER ONE:</td>
<td>15</td>
</tr>
<tr>
<td>Statistical Methods, American Society and Experts</td>
<td></td>
</tr>
<tr>
<td>CHAPTER TWO:</td>
<td>53</td>
</tr>
<tr>
<td>Psychology’s Experts: Objective Arbitrators and Managers of Methods</td>
<td></td>
</tr>
<tr>
<td>CHAPTER THREE:</td>
<td>113</td>
</tr>
<tr>
<td>The Critics of Correlation: E.G. Boring and Beardsley Ruml</td>
<td></td>
</tr>
<tr>
<td>CHAPTER FOUR:</td>
<td>164</td>
</tr>
<tr>
<td>T.L. Kelley and the Science of Mental Measurement</td>
<td></td>
</tr>
<tr>
<td>CHAPTER FIVE:</td>
<td>207</td>
</tr>
<tr>
<td>The Meanings of Measurement</td>
<td></td>
</tr>
<tr>
<td>CHAPTER SIX:</td>
<td>259</td>
</tr>
<tr>
<td>Professional Politics and Methodological Debates</td>
<td></td>
</tr>
<tr>
<td>CONCLUSION</td>
<td>316</td>
</tr>
<tr>
<td>BIBLIOGRAPHY</td>
<td>326</td>
</tr>
</tbody>
</table>
Abstract

Methodology makes visible to a scientific community the phenomena disclosed by research activity. The acceptability of methodology is inextricably tied to the acceptability of the data resulting from a particular method. Therefore it is clear that it is important not only to understand the methodology of a particular science, but of equal importance to understand the processes through which particular methodologies have become acceptable to a community of scientists. This thesis focuses on the processes through which statistical methods became acceptable to psychologists.

I identify two competing research traditions—specifically experimental and correlational psychology—and display their different interpretations and uses of statistics. I argue, however, that it would be wrong to credit the opposition between these research communities as owing to their conflicting ideas about the meaning of statistical methods. Rather, their conflict stems from differences over how science should be practiced, how labor within the research community should be organized, how knowledge should be generated, who should generate it, and who should apply it. In other words, I approach these conflicts over the interpretation (and uses) of statistics as reflecting differences in the intellectual, social and technological interests that operated within these research communities.
Acknowledgements

This thesis was written in a period of time when I changed residences three times, in three different countries. Many friends supported me, financially and emotionally, through these years. Many colleagues helped me along the way, often without their knowing it, but these people are too many to cite here. Still there are always those who seem to reach beyond what one usually expects.

John Beloff and Donald MacKenzie, who co-supervised this project, have always encouraged me and provided direction. Even though we were separated by the Atlantic, they promptly read and commented on all that I wrote. Their comments were never trivial - they have made this a better thesis. I take this opportunity to thank them for their help.

This project has been aided along through the support of great friends who listened, offered advice, and provided a rich fellowship. Richard Arnold, Michael Barfoot and James Cheesman eased the tensions of work and provided me with many memorable moments of a time that can never be repeated. Nancy Donehower has been a constant and valued friend, as well as a sounding board for some of the ideas in this thesis. I have to thank Elizabeth Hartzell, who despite a full schedule of responsibilities, has been understanding and patient with me while I travelled through the emotional pendulum that follows from working on PhD thesis.

I would like to thank the Social Sciences and Humanities Research Council of Canada for providing me with a Doctoral Fellowship for the three years I resided in Edinburgh. Also I would like to thank Edinburgh University for providing an
Overseas Research Scholarship that helped defer the cost of tuition.

The research in this thesis is based largely on archival sources. I would like to thank the Libraries and archivists at Harvard, Stanford, Johns Hopkins, Columbia and Teachers College, for allowing me access to their holdings and for helping me to locate the sources I required. I thank them for their permission to use the documents herein cited.

Over the past three years I have taught at the University of Saskatchewan. The secretaries have helped me out with typing and getting drafts into a presentable form. The Department Head, Art Clark, gave me free access to office equipment and supported this work in a variety of ways. The President’s Special Research Fund (provided by the SSHRC) at the University of Saskatchewan has supported various projects that were related to this thesis. I thank the University and the members of the Department of Psychology in particular, for their support.

At the end of all of this, I still have to thank Robert G. Weyant who, as my MSc. thesis supervisor, made me want to write history.

Outside of these considerations, I declare that this thesis has been composed by myself.
Introduction

Statistics provide what Steven Stigler has called a "quantitative technology for empirical science." The adoption of these techniques by psychology might seem to be a rather straightforward matter. If this were the case one possible scenario may suggest that as psychology became "more scientific" and developed more sophisticated procedures to quantify psychological phenomena, finding ways to analyze this data represented a logical progression. In this study of psychology's assimilation of statistical methodology I find no evidence for such a "logical progression."

Rather, I find hard fought debates that are neither won nor lost. I find tactical exchanges and political strategies intermingling with directives on methodological procedure. At times the debaters seem as confused about the issues as those who followed the debates. But it is not so much a matter of confused issues as it is a matter of issues being freshly defined.

In this work I adopt a sociology of scientific knowledge perspective. Thus I see no firm connections, no compelling deductions, that fit together a chain of research activity. By this I mean that I do not accept that a good research idea carries with it a program of implementation that in turn suggests the correct methodological procedure. Rather it is argued that chains of research activity are socially constructed.

The actual means through which scientific knowledge is socially generated is of course difficult to lay out in
detail. To ignore the social patterns of research activity, however, makes the suggestion that these are not important in understanding scientific developments. To lay out in any detail why this might be so presents an equally difficult task. Thus I have chosen to write this historical account from a sociology of knowledge perspective, rather than debate the merits of undertaking such an approach.

Exploring psychology's history from this standpoint aids our understanding of statistical methodology adoption. Without such a perspective I find it difficult to account for the readiness with which statistics were adopted by psychologists. Also by proceeding to place statistical debates within a sociology of knowledge perspective, the seriousness of the objections raised by other psychological researchers to their application becomes revealing and suggests why these debates remain salient to this day.

In the chapters that follow I draw attention to the patterns of group membership among researchers. One way I do this is through an examination of psychologists' private correspondence. The exchange of letters among researchers reveals - directly and indirectly - research networks. Through these informal exchanges one can establish who is respected in the field, who is disliked, and who is ambitious. But it is not just these interpersonal matters that come to light. The candidness of their discourse with respect to research proposals can be seen against the formal document. Also, through this patchwork of letters it becomes possible to identify those group members who exercised considerable influence within psychology. This
influence can be seen in terms of their ability to direct funding as well as membership to professional groups. In the archives we find evidence of the day to day politicking that went on within psychology.

In this thesis I draw upon these materials to demonstrate that psychologists' professional ambitions and social visions merged with their methodological practices to create competitive networks of research activity. I attempt to show that arguments over methodology are essentially arguments over who should have the right to generate and apply psychological knowledge.

The theses is organized to draw together three broad areas: the social, the mathematical/technical, and the politics of professionalism. Thus the first two chapters set the social scene in progressivist America and the development of expert systems for dealing with social problems. The next three chapters examine the technical details of methodological debates. By placing these debates in the context of the competing models of applied research, I try to show that much more than ideas were at stake in these controversies. Indeed a "life-style" of research was threatened. In the final chapter I use archival materials extensively to show the intricate manner in which professional politics interacted with methodological programs.

The first chapter begins with a more detailed account of the type of sociology of knowledge perspective I adopt in treating methodological debates. The principal aim of the chapter, however, is to draw together the broader social
issues that were faced by the community of psychologists. It becomes clear that in America science was promoted as one way to resolve social problems in an objective manner. "Objective" knowledge was taken to mean that the knowledge generating practices, and therefore their product, was insulated from the biasing effects of human agency.

It was the hope of many reformers that the social sciences would provide the experts who would redesign and engineer a new social order. Yet there was no agreement as to who was to be considered an expert. Thus, I argue that since there is no compelling deductive connections between acceptable scientific knowledge and the specific nature of applied recommendations, the psychologist-expert was defined through two opposed models. Each model proposed different strategies to remove the effects of 'human agency' in knowledge production.

The first of these 'expert models' I term the "objective arbitrator". On this model psychologists maintained an active role in the production of knowledge via their role as an arbitrator among the different facts and methods used in research. They promoted the "objectivity" of their judgment.

I call the second model of expert the "managers of methods". In this model, methods themselves are assigned meanings that imbued them with inferential capacities. Thus psychologists did not make inferences; their methods were taken as indicating the generalizability of their research findings. Objectivity was embodied in the methods, not in the "disinterested" judgments of the psychologist. The role
of the psychologist was to manage methods and thereby produce knowledge that was objective. In the remainder of the chapter I suggest that the rapid adoption of statistical methods by psychologists was due largely to the influence of the second model of "expert".

The second chapter sketches the development of these two models of expert. Each model was developed within a different institutional context. The "objective arbitrator" model was tied to the university system through the laboratory. The laboratory provided a close-contact environment and experimental technique and judgement were highly prized. In this model "craft knowledge" and the development of experimental intuition were rewarded. The environment was such that researchers were in close contact with one another and the group was well informed as to the research of their laboratory colleagues. Also, this closed system promoted an intimacy with the production of data and the nature of the data produced.

The "manager of methods" model was derived more from the corporate/management structure that was evident in the business environment. This model assigned management priorities and directed research activities. Research was made more "efficient" through a process of specialization. Methods were adopted that could be "book learned" - thus there was an increase in graduate school instruction in such methods (e.g. statistics and test development fell into this category.)

From a management perspective, the use and development of research methods was seen largely as a technical matter.
Thus training in the use of a method, such as statistics, was largely a technical training. The implication was that not all users had to possess a sophisticated understanding of these methods. Thus graduate students were used to collect and analyze data. The data generating activity as well as the data analysis were seen as strictly technological endeavours. The director of the research supervised the deployment of these methods and interpreted the findings.

Because this research was carried on outside of the university, it was important to present methods in ways that would convince other professionals of their worth. This was different from convincing one's own research colleagues. Also it was important not to run roughshod over the opinions of other professionals. Thus methods were favored which could instill agreement among a diverse selection of professionals.

I argue in this chapter that statistics provided such a method. The inferential capabilities of statistics were promoted and they were interpreted so as to remove any hint of human agency in their use. Thus these methods were believed to represent what was inherent in the data and did not reflect the opinions of the psychologist.

In the third chapter I begin by noting the popularity of statistical methods among psychologists. This attention to statistical methodology attracted criticism. These critics were largely from the experimentalist camp of psychology. I point out, however, that some of them were applied psychologists.
The chapter focuses on the works of E.G. Boring and Beardsley Ruml. I chose these researchers because both were prominent and influential. Also I considered it important that both Boring and Ruml were redressed by Truman Kelley. This allows for a natural unity within the narrative. I also chose Boring and Ruml because the former was an experimentalist and the latter an applied psychologist.

I begin with Ruml's critique, the upshot of which was that mental testers, in their willingness to use correlational techniques, produced data that was of little theoretical value. He proposed that mental testers inclination to use statistical methods when the assumptions of linearity were not met led to the production of "derivative facts". Intelligence testing had made no theoretical advance, he thought, mainly because the knowledge generated was misleading. He also objected to combining subscales on intelligence tests as these represented qualitatively different dimensions. The only warrant Ruml could find for proceeding to create such aggregates, was that mental testers were committed to the use of statistic techniques. Their motivation, he thought, was not in keeping with the spirit of science.

Boring's criticisms of the use of statistics in mental measurement were more elaborate. He set out his criticisms under three basic analytic categories: philosophical, mathematical and empirical.

His philosophical objections stemmed from his Bayesian perspective. He argued that science proceeds from "cogent reason". His criticism was that mental testers were willing
to make statistical assumptions on the grounds of "insufficient reason". The upshot of this was that he thought unless it could be experimentally demonstrated that the normal law held for mental test data, one should not assume that it is a reasonable assumption.

His second objection was mathematical. He pointed out that if the normal distribution held for one set of units, X, then it would not hold for \( X^2 \), and vice versa. His argument was that if nature conformed to the normal distribution, then the researcher must find "nature’s units". Boring argued that if nature’s mental unit is X and if we happen to use in our research a mental unit that is \( X^2 \), then the normal distribution can not possibly hold for this unit.

Boring stated that units in psychology are not "Nature’s", but are chosen arbitrarily. He reasoned that unless it could be empirically demonstrated - through laboratory studies - that the normal distribution held for a chosen unit, it was unreasonable to assume that such an arbitrary mental unit followed a Gaussian distribution.

This led into his third objection, what he called his empirical analysis. Here Boring asserted that there was plenty of evidence for non-normal distributions in "Nature". He also pointed out that, given this non-normality in Nature, it might be possible to choose units of analysis that were normally distributed. Finding a unit that was normally distributed then did not constitute evidence that the unit was "Nature’s". The meaning of the unit had to be ascertained. Thus he saw it as the job of the researcher to
determine the "psychological" meaning of a unit of measurement. This led to his argument that units of measurement in psychology were not "quantitative" and therefore measurement reflected only a rank ordering. Statistics that assumed equal intervals - such as means, standard deviations, and correlation coefficients - he argued were inappropriate for analyzing rank ordered data.

Both Ruml and Boring believed that to interpret statistical methods as being inherently inferential required the psychologist to make unjustifiable assumptions, which in turn led to unwarranted conclusions. They both argued for the priority of a more qualitative analysis of the data before proceeding with a quantitative one. In drawing out these critiques I point to the intellectual, technical and cultural dimensions that were pulled together and reflected in their criticisms. This leads to the second point of the chapter: conflicts over the uses of statistical methodology were due principally to the competition between different knowledge generating strategies (and the implied technical and cultural factors) and not to the applied/pure science distinction.

The fourth chapter lays out Kelley’s response to Boring’s and Ruml’s critiques. I begin by showing that Boring’s critique was supported by many influential psychologists. I then set out Kelley’s position of influence both in the mental testing community and as a statistician.

Kelley’s strategy was to show that the use of statistics proceeded according to acceptable scientific standards. He suggested that both Boring and Ruml were at times
inconsistent in their critiques. He presented both of their arguments as reflecting an insufficient understanding of statistical methodology. This led them, Kelley thought, to make inconsistent statements. He also argued that because statistics was a "scientific" method, the criticisms against their use were essentially criticisms against science.

Kelley and Terman wrote a response to Beardsley Ruml. Their rejoinder reveals the importance of correlational techniques in mental measurement research. In rejecting Ruml’s argument that it is inappropriate to form aggregates using subscales in an intelligence test, they utilize examples that affirm the usefulness of correlational techniques in mental measurement research. The examples and metaphors they chose reveal their commitment to interpreting correlation as a means of identifying underlying patterns in the data.

It becomes evident in this chapter that Kelley interpreted statistical methods as an inductive technology. Evidence for this is drawn from his rejection of Ruml’s argument that a definition of intelligence should be the first order of business. Ruml argued that it was impossible to measure something that was undefined. To proceed without a definition, he thought, led to a variegated assemblage of theoretically neutered facts. Kelley saw this as an insistence that science must proceed in a deductive manner. He rejected this as too restrictive and insisted that statistics provided an inductive technology that would build up the "facts" upon which a theory of intelligence could be constructed.
It was Kelley's interpretation of the capabilities of statistics that provided the impetus for his criticisms of Boring's papers. He argued that Boring thought of statistics only from a deductive perspective. He argued that this limited Boring's understanding of statistics and led him to conceive of them in a narrow fashion.

Kelley read Boring's criticisms as being inconsistent. From Kelley's perspective, any employment of the probable error involved an acceptance of the normal distribution. Thus he reasoned that Boring was inconsistent in recommending the calculation of probable errors while at the same time criticizing the normal distribution as a plausible assumption. He saw Boring as inconsistent when he suggested that statistics such as the mean, standard deviation, and correlation coefficients could not be applied to rank ordered data. Kelley saw all statistics - even those Boring recommended for rank ordered data (medians, quartiles and contingency coefficients) - as requiring for their interpretation the assumption of the normal distribution. Thus the upshot of his critique was to suggest that Boring inadvertently recommended techniques that required the very same statistical assumptions that he inveighed against.

In the fifth chapter I focus on two specific issues that divided Boring and Kelley. In this way I attempt to show that the broader social concerns outlined in the previous chapters are visible in their discourse over very specific, technical matters. The first issue deals with the interpretation of the probable error of mean differences. The second issue concerns the meaning of a "unit of
measurement".

Boring used the probable error as a descriptive statistic. Kelley interpreted the statistic in a pragmatically predictive manner. In their exchange of letters over this issue the impression is given that they did not really understand each other. The chapter outlines some background concerning different understandings of the probable error. I argue that the functional role played by the notions of variability and "error" within their distinct research communities accounts for their apparent incommensurability.

The second part of the chapter examines different perspectives as to the meaning of a "unit of measurement". I outline three approaches to measurement theory—classical, representational, and operational. Classical theory regards the unit of measurement as a quantitative portion of that which is being measured. Wundtian psychology, classical psychophysics and Titchener's structuralism worked from this understanding of measurement. For reasons I outline in the chapter, there was a drift away from classical conceptions of the measurement unit. I suggest that two opposed models were put forward, though neither was well formed until the 1930s. These models were "representational" theory and "operationalism".

Representational theory presents measurement as a procedure through which numbers are assigned so as to preserve the empirical qualitative relationships among objects. Operationalism, on the other hand, suggests that numbers assigned to objects do not have meaning outside of
the measurement operations used. Thus on the first theory of measurement, numbers represent empirical relationships. Number assignment provided a shorthand, economical description of the empirical relationships among objects. On the second perspective, numbers in themselves were regarded as meaningless. Measurement operations placed objects into relationships with each other and it was these "operationally derived" relationships that served as 'data' for theory construction.

I argue that these different theories of measurement were maintained and further developed within the experimentalist and correlationalist research communities. Thus, I argue that these measurement perspectives bring together the intellectual, technical and cultural interests of these research programs. E.G. Boring and T.L. Kelley are again used as representatives of these communities. Boring worked from a "representational" perspective while Kelley was an "operationalist".

In the final chapter I show the ways in which methodology and professional politics interacted. Because of this interaction, I argue that people were tactical as well as strategic in their dealings with each other. In other words, I draw together from archival sources details as to how the methodological/professional divide was mixed up with the rather messy business of week-to-week professional politicking.
Notes

CHAPTER ONE

Statistical Methods, American Society, and Experts

Methodology makes visible to a scientific community the phenomena disclosed by research activity. That is to say that research findings resulting from an assortment of random procedures, chosen by the whims of a researcher, would have little or no meaning to a scientific community. The way in which data is disclosed is as important as the data itself. The acceptability of methodology is inextricably tied to the acceptability of the data resulting from a particular method. Therefore it seems clear that it is not only important to understand the methodology of a particular science, but it is equally important to understand the processes through which particular methodologies have become acceptable to a community of scientists.

The processes through which a methodology becomes acceptable are seldom apparent to the historian or the scientist. As Horace Judson commented:

Once the Humpty-Dumpty of discovery is put together, all the historians and all the sociologists can’t really scramble him again - often not even the scientists who were most closely engaged, for their memories are the first to begin to be altered by the persuasiveness of the thing discovered.¹

Although Judson’s statement was directed to the discovery of the double-helix structure of DNA, it has merit with respect
to this discussion of methodology. The method of discovery
(in the case of DNA it was the use of X-ray crystallography)
is projected as not only the correct method, but the only
method that would have resulted in the discovery. Thus
those who promoted similar methods, but who had not met with
success, become honored as predecessors. Other
methodological approaches are ignored. All of this activity
obscures our understanding of how a particular method became
acceptable. Evaluating the genesis of a methodology in
light of what was produced, biases our understanding of how
scientists develop and use methods.

One such example from psychology can be found in Lee J.
Cronbach’s recollections of our methodological past. He
seems to judge past methodology in the light of present
circumstances. In his presidential address before the
American Psychological Association (1956) he commented that
because psychology is a young science it experiences a
"rapid turnover" in interests as well as "theoretical
concepts". Yet he suggested that inspite of this apparent
flux in the discipline, "our methods of inquiry have become
increasingly stable". Adding to this, he stated that it was
these methods which qualified psychologists as scientists.2

One aspect of Cronbach’s perspective that I find
disconcerting is that it suggests that historical
circumstances differentially affect theoretical issues and
methods of inquiry. That is, the first remains in flux
while the latter becomes increasingly stable. Why should
our methods of inquiry advance toward stability?

I don’t think they do become more stable. Indeed, while
Cronbach is putting the "Humpty Dumpty of discovery" together, he is confusing the appearance of stability with stability. Or, to say it in terms more akin to Cronbach's field of expertise, it is like confusing the state with the trait in personality assessment.

Method suggests stability since it suggests orderly activity. It is generally held that methods disclose the "facts" that figure in debates. In itself, method is seen merely as a means of extracting and making visible the data. Generally, so long as the facts which are generated through a certain set of procedures are accepted as facts, these methods become established as the correct procedures. It is through this association with "facts" that methodology escapes debate and therefore appears to be stable.

Matters of "fact" are held in science to be permanent. If new methods result in contradictory facts, either the theories are wrong or the "facts" are not facts and the methods associated with them are judged as procedural errors. Shapin and Schaffer (1985) have devoted lengthy discussions to such issues in connection with Hobbes's objections to Boyle's scientific experiments. They commented:

In the conventions of the intellectual world we now inhabit there is no item of knowledge so solid as a matter of fact. We may revise our ways of making sense of matters of fact and we may adjust their place in our overall maps of knowledge. Our theories, hypotheses, and our metaphysical systems may be jettisoned, but matters of fact stand
undeniable and permanent.\textsuperscript{3} Generally so long as a fact stands firm, the methods used in generating it also stand firm. Under some circumstances (such as technological advances in apparatus) methods change, but only if they do not change the facts. They also change under conditions when the facts themselves become judged as 'artifacts'.

Shapin and Schaffer suggest that the apparent stability of "facts" resides in their being viewed as owing their existence to nature rather than to human agency. Thus "facts", and the methods that make them visible, are depersonalized through what Shapin and Schaffer call "objectifying resources".\textsuperscript{4} As they so aptly put it:

In common speech, as in the philosophy of science, the solidity and permanence of matters of fact reside in the absence of human agency in their coming to be. Human agents make theories and interpretations, and human agents therefore may unmake them. But matters of fact are regarded as the very "mirror of nature."\textsuperscript{5}

Facts are dependent on procedural regularity. Thus they are tied to a body of techniques that constitute what we call "methodology". This procedural regularity results from social agreements. The interpretation of methodology, as well as the "facts" it produces, speak to - as well as reflect - the social organization of research groups.

Methodological debates can therefore be placed into two basic categories, each involving distinct social interests. That is, in some methodological debates the facts are not
seriously questioned but the means of production of these facts are. These conflicts are initiated when "better" methods appear to produce the same phenomena. Usually methodological tensions such as these arise when new technologies are introduced. Thus what we see are internecine conflicts where younger scholars challenge the merits of both the old methods and the status of the "old guard".

In the other type of methodological debate, the facts are labeled as artifacts, and the methods associated with the production of these facts are displayed as inappropriate or as misinterpretations. In these conflicts the competition between research groups is more intense as there is more at risk if one group is judged as losing the debate. Indeed the "loser" in such conflicts loses credibility in the science community.

Typically sociologists of science who locate their programs of research on scientific controversy choose to focus on this second type of methodological conflict. The first type of methodological conflict is certainly capable of sociological analysis, but appears with less frequency in the sociology of scientific knowledge literature. Certainly part of the reason for this is that in the first type of debate, the participants are less divided on their social and scientific interests. In this thesis I focus on the second form of methodological debate.

By taking the view that methodology reflects agreements among researchers as to the procedural ordering of a set of techniques, a number of questions are posed - all of which
are controversial and have met with considerable discussion. How do these agreements arise? Do scientists proceed to discover a set of techniques that can be applied in a particular order to determine the best solutions to the problems at hand? On this view, methodological agreement reflects a consensus that the outcome resulting from the use of a set of procedures provides clear solutions to research problems. This consensus also assumes that there is agreement as to the interpretation of the methods used.

Another view of how methodological agreement arises attributes more to the social and cultural aspects of research groups. Agreement as to the ordering of a set of techniques, as well as their interpretation, is seen in the context of the social relations within research communities. That is, from this perspective methodological consensus is seen not as epistemologically determined so much as it is socially constructed. Thus, agreement as to how to interpret and use a set of techniques reflects agreement on how to organize peoples research activities to practical ends. Methodological agreement is seen in relation to its role as regulating social interactions within the scientific community.

It is not controversial to admit that scientists agree and disagree on methods and the results derived from such methods. Thus in this trivial sense science is social. Nor is it any longer considered controversial to suggest that different ideologies use science, and direct science projects, to assert their special interests. When one suggests, however, that scientific knowledge is
constitutively a social product then strong objections have been raised.

These objections stem from the question as to whether or not science is essentially a passive observation of nature. On the one hand, it is argued that methods enable the scientist to observe and to make visible a natural order of things. In other words, methods extend the researchers observational capabilities. This is a popular conception among scientists, and refers to the first interpretation of methodological agreement I set out above. Understanding methodological agreement as agreement on the best ways to extend observational capability leads to the suggestion that methodology serves no social function. Research groups agree to use a particular set of methods because these provide the best ways to view nature. Agreement is determined by dimensions external to the research group. Thus the best methods provide the best "glimpses of nature."

Such arguments eschew the role of agency in science. The proponents of this "glimpse of nature" view of science, object to the notion that scientific beliefs correspond to the culture in which they exist. The architects of these critiques generally hold the view that scientific beliefs are determined by nature and identify as "relativist" any sociology that emphasizes scientific knowledge as the product of scientists' activities. They apply the label 'relativist' to anyone who regards scientific knowledge as a social construction. Furthermore, these critics generally assume that if science can be demonstrated to be constitutively social, then such science tells us nothing
about nature.

The tension flows, therefore, from those who argue that science is constrained by nature to those who detail the role of social agency in the construction of scientific knowledge. The former group perceives the latter as 'relativists' and essentially antiscientific. The latter group regards the former as mistaking scientific facts for glimpses of nature when they are really social constructions.

Such tension is avoidable and the two perspectives need not be portrayed as diametrically opposed. Indeed, Andrew Pickering (1985) has suggested that both groups, what he calls the "realists and relativists", are partially correct. He wrote:

Empirical studies certainly justify the relativist imputation of agency to scientists, but at the same time point to constraints upon that agency - to a degree of passivity in research practice. There are both active and passive elements in the evolution of scientific culture.8

He suggested that experimental practice is constrained by the need to place findings within an "already articulated field of established knowledge". This can refer to both the methodology that established the visible indicators that constitute knowledge as well as the social agreements as to the interpretations of those methods. Thus the constraint is social. As Pickering points out, any experiment (or method for that matter) exists in relation to what he calls a "range of precedents." These include agreements on what
is appropriate apparatus, what is the proper interpretation of the phenomena made visible through the inscription devices, as well as what is considered to be "benchmark" cases.

Pickering suggested another constraint on experimental practice, that of "unquestioned craft practices". He argued that experimenters are "not entirely free to generate whatever observations they chose" because they are dependent on "tacit skills, routines, habits, etc."9 Any individual scientist, therefore, is unlikely to interpret data in any individualized fashion. 'Science practice' is socially and culturally constrained.

Once it is acknowledged that methods and experiments operate within a field of socially constituted constraints, it need not follow that nature has no role to play in the construction of methods and theories. To put the matter simply, it is conceivable to argue that nature has a say in, though it does not determine, the theories and methods that operate in a science. This "say" is indirect and is manifested in a negative way. As Pickering commented, it "makes sense to speak of nature as manifesting resistance to the interpretations we place upon it".10

This is akin to J.J. Gibson's (1977) notion of "affordances".11 That is, nature affords some interpretations, but resists others. In our sciences we keep those interpretations of nature that are resisted least, but this in no way undermines the argument that these interpretations are socially constituted. However to confuse an 'affordance' of nature with what constitutes
nature, or describes nature, would be fallacious.

Although it is possible to consider theories, methodologies, experiments, etc. to be constitutively social, this need not imply that our theories, methods, etc., have no correspondence with nature. What we learn about nature through the science process is admittedly indirect, but we do learn something. Pickering summarized his argument:

The achievement and maintenance of traditions of experimental research, in which each experiment can be said to reproduce and explore the same natural phenomenon as its predecessors, does tell us something about nature and not solely about culture.12

In this thesis I am focusing my discussions on methodology, specifically statistical methods. I am not dealing with ‘discoveries’ except in such instances as when psychologist-statisticians state that they discovered a new formula. Thus I am not so concerned with the correspondence between discoveries & theories and what they tell us and do not tell us about nature. In focusing on methodology my inclination is to look at the social/cultural context that spawned psychologists’ interest in applying statistical methods to psychological problems.

As I stated in the opening paragraph, methodological agreement presupposes many dimensions that have a direct bearing on theory construction as well as what is regarded as data. Scientists use methods that they believe will provide the means through which information can be extracted
from nature. Yet within these methodological agreements one can locate a consensus on the regulation of social interactions within the scientific community. Methodology and what presumes to be analyzable data are inextricably tied to social patterns of knowledge generating activity.

From this it follows that statistics, like all human creations, have a social, political and biographical history. How statistical techniques are presented to the public, the strategies used to demonstrate the social and scientific benefits arising from statistical practices, their role in the politics of professionalism and in fulfilling personal ambitions all have a part to play in the development of statistical methodology. To view the development of the use of statistics outside of this context underestimates the ways in which personal enthusiasm, professional ambition and social vision work together with mathematical logic to create the dynamic patterns of discourse that make methodological debates possible.

At the turn of the century, psychology was beginning to challenge the nineteenth century ideals of what constituted scientific knowledge. As an academic discipline it was struggling to separate itself from philosophy. It accomplished this by emphasizing their mastery of a methodology that was different from that of the methods philosophers used. That is, they promoted a set of quantitative techniques.

Although psychologists often times lost to their philosopher colleagues the arguments over whether or not mental phenomena could be quantified, they continued to
design and interpret methods that provided "quantitative data" that supported their contentions. Even if psychologists were sometimes less than convincing in their quantitative treatments of psychological phenomena, they clearly demonstrated that they were doing something other than philosophy. They earned the attention of university administrators and received money to build facilities that supported their research.

Having gained the identity of 'not being philosophy', tensions grew within the discipline. Many different types of methods were being promoted as scientific and capable of revealing quantitative truths about mental processes. The growing enthusiasm for a methodologically pluralistic psychology was not shared by all psychologists. Some researchers believed that it was only through specific sets of procedures that a psychological quantity could be meaningfully evaluated.

Thus the first thirty years of American psychology are marked by a patchwork of methodological controversies. The social pressures of the time, as well as the growing professionalism among psychologists, fed into - and were reflected in - these methodological debates.

In the sections which follow I describe the tensions that were apparent in American society at the turn of the century. At first it may seem odd to discuss such a context. Its importance lies, however, in achieving an understanding of two peculiar aspects of science in America. The first being that American science gravitated to practical problems. Applying scientific knowledge was
clearly important. The second aspect is that American science appears at times to emphasize method over content. During the new century it was the scientific method that held great social promise for restoring order to chaos.

The Dismantling of the American Dream:

Statistics emerged as a research "tool" for psychologists during a period of time when America was going through what Henry F. May described as a "cultural revolution". The developments in American culture that gave rise to the class of "experts" also encouraged particular interpretations of "science" and "method". These interpretations had a bearing on how psychologists understood their own methods.

American society at the turn of the century was marked by social turmoil. Much of the social upheaval was owing to the rapid growth of the cities. Between 1860 and 1910 the urban population multiplied almost seven times. New facilities for transportation, policing, housing and sanitation were required at an unmanageable rate.

The immigrants contributed substantially to the growth of the cities. In 1907 1,285,000 immigrants were living in the United States. By 1910 this number had soared to an estimated 13,345,000, comprising one seventh of the national population.

This growth in population brought with it a growth of slums. Here were the violent strikes and the clubbings, the murders, the extremes of poverty and the open graft of the political bosses. Lincoln Steffens, who wrote about the New
York ghettos while stationed as a reporter on the "police beat", commented on the turmoil of Ghetto life:

The tales of the New York Ghetto were heart-breaking comedies of the tragic conflict between the old and the new, the very old and the very new; in many matters, all at once: religion, class, clothes, manners, customs, language, culture. . . . It was a revolution.16

Clearly one of the implications of Steffens' writings, as well as those of other reformers, was that the American Dream was not being realized in the immigrant classes.

The reasons offered for this failure of immigrants to partake of the good life were that they, by their nature, were unintelligent and gravitated to undemocratic ways of life. Steffens observed that reformers in New York, St. Louis, Minneapolis and other cities, were urging their State Legislatures to take away from such municipalities powers of self policing and self-government. These reformers believed this to be a first step in removing corruption. Later, in his Autobiography (1931), Steffens reiterated this sentiment:

In those days educated citizens of cities said, and I think they believed--they certainly acted upon the theory--that it was the ignorant foreign riff-raff of the big congested towns that made municipal politics so bad.17

It was thought that local sources of power in these communities corrupted municipal government. But corruption was not so easily rooted. Nor was corruption part and
parcel of minority culture. It took the exposure of the corrupt practices of the white, anglo saxon, protestant, businessman to persuade the masses that prescribing solutions to society's problems required more than short-sighted political remedies.

Muckraking America: Coping Corruption

In the latter years of the 19th and early 20th centuries, there was a tremendous growth in the trustification of industry. In a list of trusts prepared in 1904 it was shown that out of the 318 listed, 234 of them were organized since 1898 and these controlled a capitalization of over six billion dollars. Economic power was perceived to be in the hands of a few and the threat of an economic imperialism was real to some critics of the corporations. The newspapers went after big business and muckraking corporations proved to be profitable copy. Ida Turnbull, who exposed the seedy business practices of Standard Oil, remarked that the public was coming to believe that: "the inevitable result of corporate industrial management was exploitation, neglect, bullying, crushing of labor," and that the only hope was to destroy the corporate system.

The muckrakers were investigative journalists and they took as their job the exposure of corrupt practices in business, labor and government. The most visible and the most prolific of the muckrakers were Lincoln Steffens, Ida Turnbull and Ray Stannard Baker. They found corruption everywhere, in everyone. As Steffens put it:
"Who" was our question still, not yet "what," and most people today do not ask what causes crime, corruption, and war, but who is the guilty man or men.20

And the muckrakers named names and this not only sold copy, it sold the notion that humanity was basically evil, self-seeking and irresponsible. In The Shame of the Cities (1906) Steffens indited the whole of the American public:

... no one class is at fault, nor any breed, nor any particular interest or group or party. The misgovernment of the American people is misgovernment by the American people.

He commented on this statement in his Autobiography (1931): "The typical American citizen is the business man, a bad citizen. If he is a big business man, he does not neglect, he is busy with, politics, and very businesslike.21

The American dream was being dismantled. The promise of the good life was no longer the obvious consequence of good living. To the contrary, sometimes it seemed that the good life—defined in economic terms—was more the consequence of criminal practices. This attitude was pervasive. During these years, prior to and following World War I, pessimistic writers such as Ezra Pound, Theodore Dreiser, H.L. Mencken and Ernest Hemingway honed their skills. What each of these writers shared was a conviction that human life was tragic, corrupt and sordid. Walter Lippmann reflected this sort of attitude when he commented that "criminal practices" were so deep in the texture of life that "anything like a surgical cutting at evil would come close to killing the...
patient."\textsuperscript{22}

But clearly the corruption in American society had to be removed, somehow. Of the solutions offered many failed. Lawrence Goodwyn in \textit{The Populist Moment} (1978) is convincing in his argument that the populist movement failed because its reform platform gave up the hope for structural change in the democratic process. More to the point, reformers gave up their challenge to the economic traditions of American society. Their focus became more limited to getting bad people out of high places, removing obvious political graft, and in standing behind legislation that would put controls on corporate power. Goodwyn wrote:

A consensus thus came to be silently ratified: reform politics need not concern itself with structural alterations of the economic customs of the society. This conclusion, of course, had the effect of removing from mainstream reform politics the idea of people in an industrial society gaining significant degrees of autonomy in the structure of their own lives. The reform tradition of the twentieth century unconsciously defined itself within the framework of inherited power relationships. Thus the "range of political possibilities were narrowed", added Goodwyn, "not by repression, or exile, or guns, but by the simple power of the reigning new culture itself."\textsuperscript{23}

Certainly Lincoln Steffens' writings support Goodwyn's observations. Steffens renounced his activities as a Muckraker as such writings merely suggested remedies for
corrupt practices in politics, business, and labor. The writings of the muckrakers created the impression, he thought, that America's social problems could be solved without disrupting the structure of society. In his Autobiography (1931) he wrote:

It was amazing to me to hear how little the muckrakers had learned from their muckraking. No wonder our readers got only our facts and the thrills of our sensations. Most of my old friends thought just what we all thought together in the beginning. They would utter the same old cliche's: "The cure for the evils of democracy is more democracy." If I suggested that the cure for the evils of political democracy is economic democracy, they would look blank.24

Steffens saw in America's refusal to change economic structures, their unbelieving acceptance of corruption; we are all "in on the evils we abhor" was his judgment of the American people. Reform rhetoric was enough for the masses.

Thus reform politics at the turn of the century were certainly not revolutionary. Reforms worked within the system. As such, American ideals of reform were tied to the traditional economic system and so were the remedies prescribed for society's problems. It was thought that the abuse of power was the consequence of the greediness of a few.

Theodore Roosevelt and Woodrow Wilson dominated American politics from 1900-1921. Roosevelt held the Presidency from 1901 to 1909 when he named William H. Taft as his successor.
Wilson held the high office from 1913-1921. Both Roosevelt and Wilson were regarded as reform politicians. The programs they proposed operated on the notion that good people in responsible positions would bring about reform. They differed on how to enact their reforms, but it is clear that neither Wilson nor Roosevelt challenged the economic hierarchy. Reform operated within the systems that be.

The control of corporate power provides a good example of their political differences. Both Presidents realized the corporations abused their privileges. Both Presidents sought to control the corporations, but in very different ways. Roosevelt regarded the trusts as the outgrowth of efficiency and, although there were abuses of power, these could be checked by having government monitor their behavior. Wilson rejected this paternalistic role for government and sought to restore the climate of free enterprise by passing and enforcing anti-trust legislation. He sought to restore free enterprise. The small, honest, businessman honored free enterprise and allowed it to work; this was the ticket to the "New Freedom". Wilson understood the small business to operate for the community and the competition of many entrepreneurs ensured honesty through a form of natural selection. For Roosevelt, corporate power required direction by those with broader social interests. The corporations were efficient organization but he regarded them as misdirected in their social responsibility. Walter Lippmann, who supported Roosevelt's approach, was accused by one critic of looking at "carnivorous teeth" and calling them "herbivorous".
Both Presidents sought to control corporations through programs that they believed would make business more honest. Their proposals did not challenge the structure of business. Both Roosevelt and Wilson, in their own ways, believed in the ideals of the business system. They adopted many business practices but the one of greatest importance here was the use of experts—consultants who would remove inefficiency. They would use experts to direct their policies and they learned from industry the political benefits of placing policy directions on the shoulders of others.

Efficiency Experts and Social Responsibility

Efficiency became an ideal during the first decades of this century. Efficient, streamlined, directed activity ensured success. The removal of the corruption that accompanied economic interests revealed the smooth operation of an efficient system. Corporations were sensitive to the reform policies of both Presidents and were also aware of the public’s interest in reforms.

The corporations had no interest in reforming their money generating practices. Nevertheless, they addressed the concerns of reformers by using experts to insure objective policy directions in matters of social concern. Thus problems were dealt with by treating them as being essentially technical matters. "Experts" were hired to solve them. As Weinstein pointed out, such a strategy kept industrial concerns "out of politics". Everyone was willing, in principle, to accept the advice of an expert.
The idea of resolving conflict or deploying an industrial strategy by first consulting an expert had great appeal. Not only was there a sense that the solution to the problem was "scientifically" determined, but—and I think more importantly—personal bias was not part of the solution. In a time when personal and social corruption seemed rampant, Americans wanted assurance that solutions to problems were insulated from personal interests. The expert promised objectivity. The reform movement sought for scientific solutions to social problems and extolled the virtues of the apolitical scientific expert.

At first the Corporations promoted the engineer as the all-round expert. They knew this expert well as it was the engineer who enabled corporations to streamline production and, through the likes of Taylor, propose more efficient use of manpower. Engineers soon became involved in issues of social concern. Monte A. Calvert commented:

A national efficiency mania was getting under way in these years . . . Thus, conservation and efficiency engineering, and through them the idea of a larger public role for engineers, became in the years from 1906-1908 highly newsworthy.28

Engineers worked their way up through the corporate structure and became leaders. The problems they faced as managers were not problems they were trained for. They were not quick, however, to turn to the social sciences for help.29 Munsterberg was keenly aware of this and lamented that the followers of scientific management principles "have recognized the need of psychological inquiries, but have not
done anything worth mentioning to apply the results of really scientific psychology."30

On the other hand, politicians were more open to have other professions serve as their experts. Certainly in matters of economy they turned to the new breed of market experts. These economists were among the first to carry out detailed case studies in the market place under government sponsorship.31 Government officials began to look to the social sciences for expert advice. Industry followed suit.

The relationship between science and engineers was one that was understood by the public. The term 'expert' when applied to the social scientist was not so straightforward in its meaning. Specifically, the expert represented know how, but not the type gained from the trial and error of informal practice. Rather, "know how" was an outgrowth of the application of the scientific method. The expert, as Samuel Haber (1964) pointed out, "knows how and why."32 The loosely defined aspect of the term "expert" involved the question of how the expert knew how.

The relationship of science to technology was left largely unexplored in any formal sense. It was assumed that the relationship was obvious. But there existed then, as today, very different ideas as to just what science is, what constitutes scientific knowledge, and when does scientific knowledge cross the gap to inform technology.33 To corporations and government the details of the interaction between science and technology were insignificant unless they pertained to either power or economy. To the scientists, however, these were matters of great importance.

36
By controlling the discourse on what constituted scientific knowledge, as well as when this knowledge was serviceable, the science community could maintain some control over its direction and identity.

Two Models of Expert

If anything was prized in the management of American Society it was efficiency. The Corporations demonstrated how one could take the day to day chaos that was part and parcel of the market place and, through efficient management, institute a control that would turn the industry a profit. Societies managers would have to be apolitical, and be the servants of society. To ensure that the managers were objective, they would have to be scientific.

Walter Lippmann suggested in his book Drift and Mastery (1914), that the new society had succeeded in breaking away from the constraints of tradition. Although he regarded this a necessary departure, he felt that the society was now failing to deal with its new freedom. This new freedom was creating drift and in its wake was chaos. The tool to bring society under control was the scientific method. He wrote:

Rightly understood science is the culture under which people can live forward in the midst of complexity, and treat life not as something given but as something to be shaped.34

The discussions of method that emerged during the early decades of the Twentieth Century spoke to issues which were tied to the application of scientific knowledge. The social
chaos brought about by urbanization, the loss of village culture, the growth of immigrant slums, the issue of race fecundity, unions and union bashing, the visible corruption in business and politics, suggested that action was necessary or American culture would drift away from democracy. The scientific method was being promoted as a means that ensured mastery over America's social evolution.

As I mentioned earlier, the manner in which one moves from science to applied science is a journey fraught with difficulty and vociferous debate. Most of this debate happened within the scientific community as they had the most to be gained and lost depending on the outcome. Those who employed the "expert" only had to be confident of the status of the service. This confidence of course was influenced by the outcome of the debates over what constituted scientific knowledge and who could generate and apply this knowledge.35

The question of who could generate and apply scientific knowledge was an important component in methodological debates. Objectivity was required of the "expert". To social reformers this implied that personal bias was not part of the scientific solution. Thus the solution was incapable of serving selfish interests and was therefore immune to corruption. This notion of "objectivity" did not prescribe anything specific about knowledge generating practices, only that they be insulated from personal bias.

Two models of the "expert" accommodated this requirement of objectivity. The first was represented in the writings of social reformers in the early 1910s. The expert was, in
the words of Herbert Croly, "a disinterested instrument", a mind fashioned to a "complete standard of special excellence". The trained mind, the exceptional individual, the disinterested advisor, was expert at making objective judgments. This type of expert was trained to be objective and was therefore attuned to what was opinion and what was applicable scientific fact. I will refer to this as the objective arbitrator model of expert.

The second model emerged out of the concern to remove any hint of personal bias. It grew out of the former model but developed so as to depersonalize knowledge itself. This model put forward the expert as a manager of methods. The advice derived was dictated by the methods and the data, not by a reasoned analysis of the data. Methods themselves extolled the properties of judgement, not the managers of these methods.

The distinction between these models of 'expert' is central to the arguments that appear in the following chapters. Historians writing about American psychology at the turn of the century have generally given attention to the tensions arising between pure and applied psychology. I want to suggest that it was not this tension, but the one that existed between competing notions of applied psychology that invigorated methodology debates.

In shifting the focus away from the pure/applied science debate to examining tensions arising from different strategies of applying psychology, I attempt to avoid the simplification that the nature of the knowledge pursued distinguishes groups of scientists. Rather it is the manner
in which knowledge is pursued (and generated) that distinguishes research groups.

Although the orthogonality implied by suggesting two models is more apparent than real, each model reflects somewhat different strategies toward research practices. It is these different strategies that become the topic of discussion in the following chapters.

Psychology, Statistics and Experts

The use of statistics by psychologists originated in the late nineteenth century. Statistical methodology was utilized mainly by psychophysicists and, at the turn of the century, by those interested specifically in mental measurement. Without drawing up a lineage of influence, James McKeen Cattell was certainly one of the earliest American psychologists to use statistical methods in his research. Charles Spearman employed statistics in a new way to make gains for his research ideas. His techniques had a tremendous impact on psychology’s methodology. Edward L. Thorndike, Cattell’s student, promoted quantitative methods and was one of the earliest psychologists to offer a course in statistics. His approach was not so much mathematical as it was conceptual. As Truman Lee Kelley remarked on his statistical education under Thorndike: "he seemed to reach conclusions through a process less statistical than my own but more comprehensive." It was during the late teens and 1920s when textbooks concerned with statistical methodology emerged as aids to the researcher. The production of these texts served as a
bridge to other more sophisticated treatments of the subject matter. Each had a preface, often written by an Editor of the series, proclaiming the necessity of statistical fluency. In the introduction to Thurston's text the editor wrote: "It is impossible for any person to read very much of present-day educational literature with pleasure and understanding unless he is acquainted to some extent with the method and the terminology employed in conducting and presenting the results of statistical investigation."42 R.S. Woodworth, in his preface to Henry Garrett's text, stated: "Modern problems and needs are forcing statistical methods and statistical ideas more and more to the fore."43 The decade of the 1920s represented not only a swell of interest in statistical methods but also an opposition to their introduction into psychology research practice.

Statistics are generally regarded as methods that are not open to the types of controversies that fester in an emerging discipline. As methods they are either employed or not employed. Yet they were increasingly applied to problems during the 1910s and grew in popularity after the First World War. As more demand was made for applied science, psychologists increasingly turned to statistics.

As the demand increased for experts, the notion of expert became more diversified. Given that statistics were being increasingly employed at this time, it is important to examine their development and interpretation in light of these circumstances.

The two "expert" models - the "objective arbitrators" and the "managers of methods" - depict two ways of doing
research. Both accepted that psychology should be applied, but they differed as to the acceptability of applied research practices. The tension between these strategies was at a conceptual as well as an institutional level. They competed for students, institutional presence and for research money. Each interpreted methodology in a manner that embraced their professional interests. Here the politics of professionalism merged with debates over methodology.

In the chapters which follow it is my intention to demonstrate that different interpretations of the meaning of statistics accompanied different interpretations of who could be considered to be an expert. An examination of these conflicts allows us to understand the social agreements that resulted in the reification of statistical methods.
NOTES


4. In their book they suggested three technologies were established as "objectifying resources". They wrote: "We will show that the establishment of matters of fact in Boyle's experimental programme utilized three technologies: a material technology embedded in the construction and operation of the air-pump; a literary technology by means of which the phenomena produced by the pump were made known to those who were not direct witnesses; and a social technology that incorporated the conventions experimental philosophers should use in dealing with each other in considering knowledge-claims." (ibid., p. 25; see as well their discussion on p. 77)

5. Ibid., p. 23.

discussions of the "Biometrician versus Mendalian" debate as well as the debate between Pearson and Yule over the correct way to measure association ("The Politics of the Contingency Table") in Statistics in Britain, 1865-1930, Edinburgh: University of Edinburgh Press, 1981.

7. MacKenzie's discussion of Fisher & Pearson's conflict reflects this type of methodological debate. He noted that this controversy "largely lacked the 'group' structure" which was characteristics of the debates between Pearson and Yule and Pearson and Bateson. (p. 210), see Statistics in Britain, op. cit..


9. Ibid., p. 6-7.

10. Ibid., p. 9.

11. J.J. Gibson, "The Theory of Affordances", in R. Shaw and J. Bransford (editors), Perceiving, Acting and Knowing, Hillsdale, New Jersey: Erlbaum, 1977, pp. 67-82. The concept of affordance is quite simple and yet intriguing as a key in Gibson's theory of perception. The easiest way to understand affordance is through an example: "A solid horizontal surface affords support. A water surface does not. A surface of support affords resting." This example is taken from John Best, Cognitive Psychology, St. Paul, MN:
12. Ibid., p. 9.


29. As David Noble has pointed out, it was not until the 1920s that engineers began to turn to the social sciences for advice in the management of corporations. (op. cit. p. 317).


33. Barry Barnes and David Edge wrote: "Conceptions of the nature of science and technology, how they are demarcated, and how they are related to each other, are historically variable and endlessly controversial." Science in Context:

34. Walter Lippmann, Drift and Mastery, op. cit., see chapter 14.

35. Barry Barnes and David Edge suggest that an expert is a "representative of a trusted institution." But, they note, that this expert must "exploit this trust" by arguing that what they say is "what science says" is true. Thus the status of a science has to be socially secure or the expert is undermined. See Barnes & Edge, "Science as Expertise," in Science in Context (Milton Keynes: The Open University Press, 1982) 233-249, p. 234.


upon the "pure/applied" dichotomy in "The Professionalization of American Psychology," Journal of the History of the Behavioral Sciences, (1973) 7 66-75. My point is simply that much of this work can be read as a history of the competition among different strategies of applied research. Most American psychologists accepted that their research should speak to practical interests. They interpreted what they were doing in research from this perspective. Thus someone like E.G. Boring, who is regarded by O'Donnell as a stalwart defender of pure psychology, writes to many of his friends about applied research in an approving manner. Indeed he himself conducted studies that certainly could be regarded as being "applied" research. In particular his work on "speed" in mental testing. Yet his strategy of just how psychology should proceed in working out an applied program differed greatly from that of the mental testing community in general. Thus the tension was not between applied and pure research so much as it was between different ways of conceiving what was an appropriate way to generate applied knowledge. Thus even though the actors themselves alluded to the 'pure/applied' distinction, in essence they were quarreling about how to proceed in generating knowledge that would lead to application. In placing the issue as a disagreement over strategies of applied research, I think the relationship between science and technology is placed in a broader context. This in turn invites interpretations of the science/technology relationship that places the "received view" of this relationship (that "pure research leads to applied
research") as a strategy of applied research which is employed to realize certain professional interests.

38. Stephen Stigler give an account of the use of statistics made by Fechner and Ebbinghaus. See his chapter "Psychophysics as a Counterpoint" in The History of Statistics: The Measurement of Uncertainty Before 1900, 239-261. (Cambridge: The Belkap Press, 1986.) His discussion of the different ways in which these early psychologists used statistics is informative. The influence of both Ebbinghaus and Fechner’s uses of statistics is difficult to assess. Although Stigler implies that their influence was profound in demonstrating the availability of these methods for psychological research, few psychologists mention their impact on this area of psychology. Indeed, F.M. Urban was critical of Fechner’s use of statistics and noted that Fechner was not very talented as a mathematician and that this might account for the lack of sophistication in Fechner’s employment of statistics. Urban wrote to Boring: "Fechner must have begun to doubt Quetelet’s view sometimes [sic] in the seventies. He collected a large material which must have convinced every sane person that Quetelet was wrong. When it comes to drawing the conclusion that the formula must be discarded, you can see Fechner trying to get up his courage and miserably failing after a time. The consequence is that he does not come anywhere and that his method for the treatment of statistical data is nothing but a poor makeshift." In the same letter Urban stated that Fechner "was not a very skillful mathematician, for he usually turned to one of his mathematical colleagues when he
had to work out a problem." (Urban to Boring, August 12, 1920. Boring Correspondence, Harvard University Archives) James McKeen Cattell who introduced the term "mental test" and conducted a series of highly influential studies in psychophysics, never mentions the influence of Fechner on his statistical practices. In particular see his letter to Helen Walker where he discussed those who influenced his statistical ideas - Fechner was not among them. (See Cattell to Walker, "Memorandum for Miss Helen Walker", Box 70, Columbia University Archives, undated. The memo was probably written as a response to Walker's letter which she sent out when she was collecting materials for the above mentioned book concerning those who influenced psychologists statistical ideas. Cattell's memo was probably written in 1926. I would like to thank Michael Sokal for forwarding a copy of this document to me.)

39. Helen M. Walker in Studies in the History of Statistical Method (Baltimore: The Williams & Wilkins Company, 1929) remarked: "There appears to be general agreement that Cattell's teaching, both at the University of Pennsylvania and, more especially later at Columbia, combined with his use of statistics in his own writings, was the greatest single factor making for the adoption of statistical methods by American psychologists." (p. 152)


Carl Seashore in his presidential address to the American Psychological Association (1912) suggested that: "Historians of the future will probably characterize the period as one of the rise of applied psychological sciences."¹ Certainly Joseph Jastrow agreed when he wrote in his autobiography that:

To speak of the renaissance of psychology, especially in the American setting, without explicit recognition of the practical motive would be a glaring omission; for the renaissance found its momentum in the appeal to psychology for the regulation of human affairs.²

Munsterberg, who early in his career cautioned against the application of psychology, had by the 1910s come to believe that the laboratory had produced enough facts to lend practical assistance to the legal, medical, and teaching professions as well as industry.³ In 1913, Seashore referred to the educational psychologist as an "educational efficiency engineer".⁴ In 1916 Walter Dill Scott was offered the first professorship in Applied Psychology and one year later the Journal of Applied Psychology was founded.⁵ Elliott Frost described psychologists' meetings as being "veritibly centered about applied interests", and it was clear that the majority of his colleagues were moving toward applied research.⁶ Testifying to this, statistical
accounts of the number of psychologists involved in applied research ranged from between 49% to 50.5% of the APA membership.7

As was evident from the previous chapter, it is not surprising that there would be growth in the area of applied research. That the architects of American society required experts in the social sciences was expected, even if the use of these experts was as much to keep issues out of politics as to gain direction. The social definition of what constituted an expert, however, was not fixed. Nor were there in psychology explicit debates over the definition of who could be considered to be an expert. Rather the question of who could be considered an expert was implicit in the different strategies for generating knowledge among competing groups.

Psychology in the Service of Society: Two Models of Expert

In the previous chapter I suggested that two models of expert existed during the second decade of the century. The first I referred to as that of the "objective arbitrator". The expert could make an evaluation and a judgement that was not tempered by personal bias. The second model I referred to as the "managers of methods". This expert placed more inferential power on the methods, and less on their own personal judgement. In this way the expert remained unbiased and objective.

Specialized books to promote applied research began to be published during the teens and twenties. There was as
well a deluge of articles on applied topics appearing in the journals. The textbooks for beginning psychologists, however, made only cursory reference to the applications of psychology. Hollingworth and Poffenberger justified the writing of their text *Applied Psychology* (1917) saying:

> There exists no book which well serves as a general text of applied psychology, presenting its principal aims, types, methods, its various fields of endeavor, and its outstanding results and accomplishments. Students of applied psychology must at present be referred to a very scattered series of special articles, monographs or books of varying value, . . . The general text books of psychology do not have the practical point of view for which he is in search.8

At first it appeared surprising that the burst of applied research activity was not carried over into the psychology textbooks of the period. But on further consideration, this seemed reasonable. Textbooks are written for pedagogical purposes. They present data that fit with theory and provide what Kuhn would call "exemplars" for a research area.9 What was implicit in the omission of extended discussions of applied research was the notion that the psychologist must master the "facts" of psychology before considering how these facts could be applied. No matter who the author was, the psychology textbooks of this period showed remarkable agreement in their tables of content. The fledgling psychologist was presented with what appeared to be an agreed upon body of "facts".10
Whether the textbook author was an applied psychologist or an experimental psychologist did not influence the content to any great degree. At first this appears odd. But to pure and applied psychologists the notion that psychology was grounded in fact—that it was a science with methods, apparatus, laboratory techniques, and tradition—was clearly important to both groups. It was commonly held that application follows from science, and if there is no science, there can be no body of laws and principles to apply. Applied psychology was seen as an extension of scientific psychology and so too would be the textbooks.

The upshot was that a student had to first become an experimental/pure psychologist—pursue knowledge for its own sake—before s/he could explore psychology's practical realm. This was the message in both the general introductory textbooks and the more specialized books on psychology's application. The first editorial that appeared in the first volume of The Journal of Educational Psychology reflected just this attitude. The Editors wrote:

Not many years ago, Professor Royce argued with force and insight the need of a "middleman" whose task should be to mediate between the science of psychology and the art of teaching—a man of careful laboratory training and with an abiding interest in the problems of the classroom. The editors of this journal believe that there is an equal need of a "middle magazine"—of a journal that shall afford a common meeting ground for the psychologist and the educator.
In the same volume Edmund C. Sanford remarked that it was satisfying to see the "coming out of this first daughter of experimental psychology." Hugo Munsterberg wrote with characteristic aplomb of the practical value of the modern laboratory. In *Psychology and Industrial Efficiency* (1913) he claimed to have sketched "the outlines of a new science which would intermediate between the modern laboratory psychology and the problems of economics." The "psychological experiment" was to be placed at the "service of commerce and industry."14

Although concern was expressed by some that psychology was not yet ready for application, there was an enthusiasm to make laboratory research serviceable. Hugo Munsterberg remarked:

> It is never a gain when a science begins too early to look aside to practical needs. The longer a discipline can develop itself under the single influence, the search for pure truth, the more solid will be its foundations. But now experimental psychology has reached a stage at which it seems natural and sound to give attention also to its possible service for the practical needs of life.15

Jastrow, another promoter of the practical ideal, introduced a series of books on psychology's usefulness by saying that "science does well to utilize the actual interests of men to build upon them the knowledge that makes for power." Science, he reasoned, should "supply the foundation in principle for the guidance of practice."16

Early in the century and certainly during the 1910's the
dominant conception of the expert within psychology was someone who could "intermediate" between the laboratory and the needs of society. The expert psychologist then, was someone who was familiar with laboratory methods. What this implied was an expert which I would describe as an "objective arbitrator". I will return to justify my description in a moment.

By 1914 another type of expert was emerging in psychology. This expert was modeled on the scientific managers who were beginning to have an effect on industry. Their methods were problem centered. Their approach demonstrated that the 'applicable fact' need not always be derived from knowledge that was obtained for its own sake. The methods were not laboratory based but were offered up as "scientific" in that they were based on observation. These psychologists spoke of improving methods that could be applied directly to problems. Although laboratory methods were scientific, and were therefore acceptable to this group, they believed that there were other scientific methods that were better suited to applied research.17

Edward Lee Thorndike, a pioneer of applied psychology, is an early representative of this type of psychologist. He conceived of the work place as his research space. In 1910 he talked of the classroom as "a vast laboratory" in which there were "many thousands of experiments of the utmost interest to 'pure' psychology."18 Here we see that the problems for research were not derived from the concerns of the laboratory - nor were the 'applicable facts'. Rather, the classroom as a laboratory offered 'pure' psychology

58
something to work on. Also, this new laboratory required methods that were quite different from the methods used in the traditional lab. Thorndike was one of the first to promote the wide use of statistical methods. These methods were well suited to the applied setting.19

Thorndike’s student, Truman L. Kelley, became a leading statistician in American psychology. Kelley saw statistical methods as leading to new insights; he did not attempt to apply laboratory facts to classroom problems. In 1914 he wrote:

The movement for vocational guidance is in its infancy, but it only depends upon improved methods and more extended research to give it a place with the older professions.

In this work he used partial correlation and the "regression equation method" because of their "peculiar adaptability" to problems in the area and this "ensured their extended use in the future."20 The focus was on the method, the statistical method in this case.

Mental testers as a group tended to emphasize statistical methods. For example, Lewis Terman wrote in the introduction to Ben Wood’s book Measurement in Higher Education (1923) that:

The reader untrained in statistical method may at first be somewhat mystified by the frequent use of mathematical terms and by the occasional reference to statistical formulae and procedures. He can be assured, however, that the essential principles involved are really very simple, at least as far as
their practical bearings are concerned. A careful reading of the text need not leave any intelligent person, however innocent of the statistical procedure he may be, greatly in the dark with reference to the real significance of facts presented."21

The implication of Terman's point was that although Wood's research was of the most thorough kind, it was readable and could be understood in terms of its practical reference. Furthermore, the details of the methods could be overlooked if one was not trained; if one was trained in statistical methodology, Terman thought that the reader would benefit even more from Wood's work.

"Facts" were no longer the sole dominion of the laboratory. Other methods were being explored and offered up as valid scientific methods. These applied researchers introduced a broader perspective as to what constituted research and, for a time, they may have dominated psychology, particularly after World War I. But with the rise of behaviorism (which conceived of laboratory methods in terms of providing a more efficient route to applications), along with the advent of experimentally devised treatment and control groups, the dominance of laboratory research was reasserted for the 1930's & '40's.22

General discussions of applied psychology during the early 1910's never mentioned the different strategies of applying knowledge and focused instead on the question as to whether or not psychology should be applied. Just prior to World War I, and certainly afterwards, discussions of
applied psychology focused on the scientific acceptability of the knowledge generated via applied methods. Applied psychologists often spoke of the scientific legitimacy of their methods, pure psychologists pointed to their scientific poverty. Laboratory methods themselves were not the object of attack.

Yet the two conceptions of applied psychology - that of applying laboratory research and that of deriving facts from field research - embraced the two conceptions of "expert" addressed in the previous chapter. The expert as an "objective arbitrator" applying facts derived from experimental research was attractive to the "pure" researchers. The expert as a "manager of methods" was more fitting with the applied psychologists who conducted research in the field.

The debate between "pure and applied" research functioned on more than one level. Not only were knowledge generating practices being debated but so too were the questions of who was to generate the knowledge as well as who was to apply it. To some degree, the reason for the different appeal of the two expert models rested with the social circumstances under which knowledge generating practices operated.

In the next two sections I outline the circumstances which favored the adoption of one expert model over the other. As I mentioned in the first chapter, the two models are not orthogonal so overlap exists. The distinction between the two models is one of emphasis. In the "objective arbitrator" model, the emphasis is on the
judgement of the scientist/expert. Although methods are an integral part of their knowledge generating routine, the inferential power rests with the judgement of the scientist/expert. In the "managers of methods" model the inferential power rests mainly with the methods. These experts promoted their methods as arbitrators and themselves as skilled technicians.

In the first section I examine laboratory based psychology in terms of its location in the university. The laboratory was an institutional presence within the university structure. I draw on this to persuade the reader that the issues facing universities, particularly in terms of the debates over the value of an undergraduate education, affected how laboratory science was conceived by psychologists. The fact that the laboratory was part of the university, that the work place of the psychologist was in the institution, made psychologists responsive to issues that concerned the university. Also, it cannot be overlooked that it was the university that provided financial support for the laboratory. Thus, psychologists who supported a laboratory approach to science tended to be sensitive to the concerns of the university.

Also, it is important to realize that the laboratory presents a close-knit collection of common interests. The labor of investigation is generally shared and contributed to by most working in the laboratory. The critical audience, those watching the methods of experiment and the recording of data, are all somewhat invested in the result of the study. As the Laboratory gains in prestige so do the
research workers.\textsuperscript{24} Thus in the first section I draw attention to the influence such a social network might have on the appeal of a particular model of expert.

In the second section I examine the growing trend in the 1910s to conduct research outside of the boundaries of the laboratory. Although this development can be seen in various branches of applied research, educational psychologists were the most effective and had the most profound impact on psychology's methodological development.\textsuperscript{25}

Applied psychological research in industry was having a difficult time during the middle and late 1920s. The Scott Company floundered after Walter Dill Scott left to become president of Northwestern University and, as Michael Sokal (1981) reported, the Psychological Corporation nearly failed during the 1920s.\textsuperscript{26} Educational psychology was successful during these years and many psychologists were employed as educational psychologists. During World War I psychologists such as E.L. Thorndike, Lewis Terman, Truman Kelley, Walter Bingham, and G.M. Whipple - all of whom were involved in education programs at one time or another - refined and developed mental testing procedures. Thus, in the second section, I focus on educational psychology.

In the situation where the research operates in the "field" (as was the case with most educational psychologists), the social relations appear to be somewhat different. First off, the research is usually conducted on a large number of subjects at one time. Thus, in data collection, the relation is that of the researcher to a
group. Also, field research would generally not involve close contact between researchers while they collected and tabulated data. Since field research does not provide a closed social environment, the methods used in collecting the data were available for criticism by non-psychologists. The researcher(s) have to justify their methods to teachers, managers, etc., who may have little understanding of the methods involved in the study. This encouraged a discourse about methodology that would be different from that among informed researchers.

Psychology in the Laboratory.

The growth of laboratory psychology emerged in America at a time when there were concerns being voiced about the quality of the scientifically trained mind. Because German science had such impact, and was conceived by many Americans as being so narrow, it was often the object of attack. Research papers, as opposed to the scholarly comprehensive reviews, were becoming the standard through which the quality of an academic was being assessed. Critics referred to the Ph.D. as too specialized and required the intellect to focus too narrowly. It was being suggested by such critics that the truly liberal mind abhorred the details of a research paper or the specialized concentration required to manipulate and utilize apparatus.

This conflict, represented as that between the Humanities and the Sciences, was most intense during the late nineteenth century and the early years of this century. By 1910 Edwin Slosson described the conflict as one that was
nearing resolution. In part this was due to a tendency in the humanities, as in the sciences, to move toward specialization. This provided not only a tolerance for, but an understanding of, the benefits accrued from the type of focused activity that was so visible in the sciences. Still, even though the trend toward specialization was shared, the sciences were the masters of the minute and the humanities saw in such detailed analysis a poverty of content. Whatever the case, the tension between the sciences and humanities was less intense and this growing tolerance was not the result of any assertion of dominance of one discipline over the other. Rather it became the case that the Professors in both camps gained something of import through the conflict.

The professors in the humanities demonstrated their worth by highlighting the inability of science to deal effectively with certain questions. They showed that certain presuppositions made by science could not be substantiated. Indeed, they showed that science was in the end subservient to philosophy. The less scientists thought about their assumptions, the more philistine they appeared.

On the other hand, the professors of science had demonstrated that what they said could at times be verified; that they were not lost to endless speculation. If they could not always provide a disclosure of the laws of nature, at least they formulated rules that allowed for the prediction of certain events. The virtues of science were often over extended by enthusiasts, but the optimism was contagious.
Mostly, it became apparent to both groups that they worked toward similar ends: that education was to broaden our understanding of life in general. In the end it was their agreement on the overriding goals of a university education that introduced some harmony to the dissenting factions. The psychologist Robert MacDougall, while promoting the importance of research, encouraged the undergraduate curriculum to include courses in literature, philosophy and history. The graduate, he reasoned, must be cultured and capable of intelligent criticism.  

Alexander Meiklejohn wrote that the American College should plunge every freshman "into the problems of philosophy, into the difficulties and perplexities about our institutions, into the scientific accounts of the world especially as they bear on human life, into the portrayals of human experience which are given by the masters of literature . . . \(^{31}\) There was a humanistic and a philosophical aspect to science as there was a science in the humanities.

Nicholas Murray Butler (later the president of Columbia University) reflected just this attitude when he wrote in the introduction to Royce's text \textit{Outlines of Psychology} (1903) that psychology, in taking up "something of the aspect of a natural science", had lost its "clearness and cogency to the philosophically minded student." He reasoned that natural sciences were based on an "elaborate series of presuppositions", none of which were "tested or examined." \(^{32}\) It was these presuppositions that clouded the interpretation of the new scientific psychology. It was the philosopher Royce who was being promoted as someone who could bring
clarity to the goings on in scientific psychology.

Royce was trained as a philosopher but had developed an interest in the "new" scientific psychology. Although he did not actually conduct an experiment in psychology, he saw such data as useful in articulating a philosophy of mind. Still, it was not methods that would extend knowledge, but clear thinking about these methods of science. In his address to graduate students at a Harvard commencement in 1910, Royce encouraged the graduates to rise above their subservience to science and become "conscious of the methods" used to pursue knowledge in their "technical branch of learning." What he suggested was that scientists had to do more than master the technicalities of their discipline. They had to represent the virtues of an enlightened mind. Scientists were challenged to show the qualitative aspects of their work.

Experimental, laboratory based psychology was seen to ensure objectivity and, as a consequence of this, provide the means of uniting all of psychology. Classic textbooks of the period pointed to the experimental method as not only the best way to generate exact knowledge, but as a means of removing a researcher's speculative bias. T.H. Ribot wrote that German psychology had set out the course that all should follow in their method of science. By turning to the laboratory, Ribot believed that psychology would "carry less and less the imprint of one man or one race." Oswald Kulpe wrote in his Outlines of Psychology (1895) that the experiment had created a "community of psychological work". Now it was possible, he thought, for every psychologist to
"enter into the methods and results of his colleagues, confirming or correcting" each investigation. He argued that soon psychologists could give up speaking of "individual systems" and speak of "psychology as a science resting on firm foundations, whose superstructure is so planned that the new fits in easily and harmoniously with what is already established."35

Experiment was thought to provide the basis for unifying all research work in psychology. Even Josiah Royce could not resist the optimism engendered by the promise of the experimental method. He wrote that "[a] centrally important modern method, which unites or may unite features belonging to all the foregoing methods, is the method of the psychological experiment."36 The methods of experimentation were appealing but their execution was problematic.

It was here that experimental psychologists found room for promoting the more qualitative aspects of their science. E.W. Scripture wrote in his classic text The New Psychology (1897) that the "sources of error in experimenting are so manifold and insidious that their avoidance and elimination has become an art which can be learned only by a specialist." He was referring to such errors as estimating the degree of "attention" the subject placed on the task, or the "pre-disposition" of the subject. Other sources of error were the factors of fatigue, practice, the nature of the apparatus, the choice of a measurement unit or errors in the definition of the problems investigated. Scripture concluded that "Only the thoroughly trained investigator can be decently certain that he has not committed every one of
them [errors] in such high degree as to make his results worthless."  

The experimental psychologist was not trained in just methods but in the judgement of the results obtained in the experiment.

It was a scientist's judgement, as an objective arbitrator, that elevated the knowledge generated through its methods. The scientist had the final say in what constituted data. Sometimes the scientist would justify his/her judgement by pointing to the reliability of the methods, sometimes by a careful reasoning based on the results of an experiment. Still, generalization was not the property of method, but a property of a scientist. Thus good scientists were exalted, not merely their experimental methods.

Models of Erudition: Scientists as Objective Arbitrators

At the turn of the century scientific research was learned as a craft under the direction of someone who was considered a master scientist. Working in a laboratory was a training ground, an apprenticeship program. The young scientist was observed, cared for, commented upon in personal correspondence. If the young scientist demonstrated that s/he had learned good judgement as well as good technique, they were placed in a job. This concentration on the director scientist no doubt was a carry over from Germany.  

The role of the eminent scientist was to teach others his/her craft while carrying out a research program. For instance, Cattell commented that American colleges would do well to learn from success. He pointed
out:

The ideal is the zoological hall of the old Harvard, where apprentices of a great man and a great teacher lived together. . . . The number of men of distinction given to the world from this small Agassiz group is truly remarkable.39

In recommending what students should look for when they go to college, Slosson - a chemist - upheld this notion that the best education is attained by studying with the best researchers. He stated:

The graduate's trend of thought and life work are largely determined by his research professor, the man who sets his course for his first voyage in the unknown. If the young man realized how much depended upon the personality and perspicacity of this pilot, he would take more pains in the selection. As it is he is apt to choose his research professor as carelessly and unpremeditatedly as he chooses a wife.

Slosson suggested further that if students would go to the best research professor, if they would "flock from all parts to a man of superior attainments wherever he might be" then not only would lesser universities get good students, but the honor of the professor would be "enhanced in his own country" and the importance of his work would be "better recognized elsewhere."40

Often in the appraisals of the first generation of American psychologists (who were often German trained) comment was made on the breadth of knowledge possessed by
these individuals. Boring wrote with admiration of Titchener's thoroughness as a scientist, his "belief in the historical orientation", and his commitment to "write English well." In his autobiography, Boring wrote of Titchener's superior and well-rounded intellect:

Titchener loved to solve puzzles, and his skill in numismatics was developed over the problems posed by Mohammedan coins. To obtain this skill he had to learn some Arabic, but he was competent with languages, and could ad lib in Latin when the occasion required it. Walter Pillsbury admired his mentor, H.K. Wolfe (who also studied in Germany). He was a "fairly close follower of Wundt", Pillsbury commented, adding that Wolfe was also "as advanced in his interests in politics as in philosophy or religion." The point of this discussion is to suggest that the experimental psychologist, because laboratories were attached to Universities, took on many of the virtues of a University professor. At the turn of the century, the debates over the quality of knowledge generated by science and the humanities, were formative on what characteristics were expected in a university professor. Certainly high on the list was a well-rounded, broad, well-informed intellect. This was a virtue that had to be possessed by both the scientist and the professor in the humanities.

In the laboratory, the student of experimental psychology was taught not only the methods of research but also the importance of sound judgement. It was as important
to learn how to think like an experimentalist as it was to learn how to manipulate and build apparatus. The student of scientific psychology was urged to take a standpoint from which to view science and was instructed not to confuse this perspective with that given by commonsense. Commonsense was value-laden; scientific facts were value free. The student must judge the facts in their own right. Titchener wrote to the beginning student that "ordinary living is not scientific". The student must unlearn the biases brought into science by commonsense. Where commonsense clearly intruded was in choosing scientific facts in light of a practical ideal. Titchener wrote:

The laws of psychology may be put to very many uses, in business, in education, in legal procedure, in medicine, in the ministrations of religion; but such uses are, from the psychologist’s point of view, by-products of his science. These practical results may be immensely important for everyday life; but science, in its impersonal and disinterested search for facts, makes no difference between one fact and another.44

As Titchener would have it, for the science to be value-free, if facts were to have a free hand in determining theory, then the scientist must be value free.

Such comments were common in early textbooks that promoted psychology as a science. If facts were freely presented and not chosen in light of practical concerns or for personal advancement, then all psychologists could - if they were disinterested in the execution of the experiment-
produce the same facts in their laboratories. This allowed for a community of science, a situation where one laboratory could check out the "facts" produced by another laboratory. Built into their notion of a "fact" was the idea of replicability. To get at the facts, to generate factual knowledge, the student-scientist must first unlearn the bias of his commonsense. Once this was accomplished, the student could judge the facts in a value-free manner and build upon this knowledge a psychology.

Psychology was a life’s work, as were all sciences, or so it was assumed. A clear implication from Cattell’s exaltation of Harvard’s Zoological hall was that for science to be successful in both research and in producing competent scientists for the future, scientists must maintain close communal contact. What was learned by the laboratory student was how to think like a scientist and this was a slow, arduous training.

The training was not just an indulgence in technique, but involved taking on a culture. The student was expected to live in the laboratory and to take an active interest in the work of his/her colleagues. Boring commented on his training in Titchener’s laboratory:

In the Cornell laboratory, . . . I became a member of the community, Titchener’s in-group. There were many of us, men and women, and we lived in the laboratory from 8 A.M. to midnight.45 Boring expected no less from the workers in his own laboratory, first at Clark and later at Harvard. In speaking of his work at Clark he stated: "We all lived
together, working the eighty-hour week in the laboratory, lunching together, being democratic." And they were productive. In three years Boring's lab produced thirty-six published papers.

This of course was the ideal. Work long hours, discuss your research, and live as a community. It was an ideal that was not realized everywhere, but in reading psychologists' recollections of their early years it's apparent that many sought after the scientific fellowship of laboratory life. Later on Boring wrote in disapproval that graduate students could be "trained" in experimental psychology through taking courses. "The way to get a Ph.D., a good one," wrote Boring, "is to live a life of scholarship and research for three or four years under the conditions most likely to stimulate intellectual development toward a prescribed maturity." In Boring's curriculum the graduate learned by "soaking up attitudes about experimental research" provided by "an atmosphere with which he [the student] is identified."46

Bruno Latour and Steven Woolgar suggested, based on their study of laboratory science, that the social situation and the professional demands are such that the "credibility of the proposal and the proposer are identical".47 In early laboratory psychology this was certainly true. The scientist to be credible and respected had to demonstrate good judgement, good technique and extol the virtues of a liberal intellect.

If applied knowledge was to be generated from laboratory psychology, the "expert" had to be an erudite, disinterested
assessor of the facts derived from research. The application of false facts would result in ineffective programs. The expert must be qualified to judge what is and what is not an applicable psychological fact as well as what is and what is not a verifiable psychological principle. Only the scholar, trained both in method and scientific judgement, could fit the bill as an expert.

Not all psychologists agreed with this assessment. As the century passed the mid-teens, a group of researchers began to force the argument over whether or not laboratory/experimental methods were the only scientific methods suitable for psychology. These psychologists sought for both methods and a professional identity that could be used outside of the laboratory.

Beyond the Laboratory: Educational Psychology and Field Study.

The shift in some quarters to abandon the psychological laboratory and to search for a new means of generating psychological knowledge is difficult to assess. Clearly, those interested in the shift away from the laboratory were applied psychologists, and certainly the most successful were the educational psychologists. What led to their belief that they could devise new ways of disclosing psychological facts is not at all apparent, though clues exist.

The first movements away from the laboratory were not in the interests of establishing a new venture in methodology. Rather, psychologists attempted to transport the experiment
to the applied setting. The means through which this was accomplished was by simplifying what was considered to be data. Introspection was dropped as it was too demanding on the subject. Instead the focus was placed on reaction time measures, number of words recalled, frequency of correct responses, etc. In speaking of Cattell’s work (one of the first psychologists to promote applied psychology), Michael Sokal (1982) commented that the tendency to oversimplify the European model of science was quite common. He wrote:

... by adopting only the mechanics of Wundt’s procedures while ignoring his broader concerns, Cattell was acting no differently with respect to his German teacher than did the American historians who studied with Ranke or the American chemists who studied with Liebig.48

The concentration became focused on what could be measured—or what could be portrayed as measurement—through simplified numerical summaries. Margaret Floy Washburn referred to Cattell’s approach as "his objective version of the Leipzig doctrine".49

Cattell, who coined the term ‘mental test’, was a leader in promoting applied psychology. His approach to mental measurement, however, was soon replaced. There was a growing trend toward the use of a Binet-style of mental testing. Another event dulled Cattell’s influence in the development of applied methods. He lost his academic position at Columbia, due in part to his opposition to America entering the War. Nevertheless, the spirit of his approach was carried on by his students. Of his students,
perhaps Edward Lee Thorndike was the most influential and industrious promoter of applied psychology.

Whether or not it was through the direct influence of Cattell is difficult to assess, but Thorndike certainly emphasized the quantitative over the qualitative in his research. In his Ph.D. (Cattell was his supervisor) he focused on the problem solving activities of cats in a puzzle box. In an analysis of behaviourism, Brian Mackenzie suggested that for Thorndike, "whatever cognitions may be said to function in the cat's behaviour" were "inseparable from the particular responses which the cat acquired in a given situation." What operated was not an association of ideas, but an association of behaviors specific to an environment. Thorndike speculated on the mental life of cats, but only with reference to the exhibited behaviors. To speculate on mental events without reference to behaviors that were functional to the organism's environment was, according to Thorndike, misguided inference.

Thorndike was never tied to animal work, and in his autobiography he was candid about his research on animal learning. He wrote that he conducted the research because of the forces of circumstance and he needed to complete course requirements. That he did not persist with animal experiments and turned to problems in more applied fields is not surprising. Thorndike and Woodworth's (1908) research on transfer of learning (the first studies to use control groups) reflected what he learned from his animal research. That is, learned behaviors are specific to a particular context, and to assume that learning Latin will benefit
one's learning of mathematics could not be supported. He opposed the school curriculum and urged superintendents to adopt a more scientific approach when setting up a program of courses. In "Educational Diagnosis", an article which appeared in Science (1913), Thorndike wrote:

Experiments measuring the effects of school subjects and methods seem pedantic and inhuman beside the spontaneous tact and insight of the gifted teacher. But his personal work is confined by time and space to reach only a few; their results join the free common fund of science which increases the more, the more it is used, and lives forever.52

Science, he believed, would revolutionize teaching and education. He was going to teach the teacher the value of science applied to their craft.

Thorndike approached problems in psychology by basing his solutions on what was observable. For him 'observable' had a very clear meaning—what could be observed could be counted, timed, or indexed in some mechanical fashion. He wrote in 1921 that: "Whatever exists, exists in some amount." His insistence on quantifying everything in a problem became so much a part of his approach that in a tribute to him on his retirement, the President of Teachers College wrote that: "To him [Thorndike] qualitative difference was merely a quantitative difference that man had not yet learned to measure."45

Thorndike's interpretation of qualitative differences as unrealized quantities, encouraged a philosophy of measurement that played down the qualitative considerations
that operate in any measuring process. For Thorndike, the point of scientific training was to learn how to make the invisible visible through measurement. Once something was visible, it could be counted. Once the problem was given physical dimensions, speculation had a basis in what was observable.

With the emphasis on the number or quantity, rather than on the introspection (through which the psychologist would seek to account for the value of the quantity), the source for making a scientific inference shifted. The discussions of the experimental subject, the reflections on experience, were now less important and were therefore not grounds for making inferences to general psychological principles. That which could be assessed as a quantity, that which 'afforded' simple numerical assignment, formed the basis of scientific generalization.

Thorndike was not exceptional in opting for a simplified approach to the philosophy of measurement. Nor were other psychologists necessarily more sophisticated in their measurement assumptions. Some, however, were more careful about the meaning of measurement. For instance, Lightner Witmer used many laboratory tests for mental measurement, but it was not the measurements per se that were important to him. Rather, of interest was the changes in mental performance on these tests after clinical intervention.46 Thorndike tallied and tabled his results and based inferences on accumulated quantities and the accompanying statistical summaries.

Thorndike was not a statistician. He admitted this.
Yet he still boasted that without mathematical training, he "managed to learn the essentials of statistical method somehow" and to such an extent that he "handled some fairly intricate quantitative problems" with very few errors.⁴⁷ He studied with Franz Boas while he was doing his Ph.D. and credited him as the one who offered him the most in terms of his statistical understanding. Thorndike was committed to the use of statistical summaries and it became a trademark of his teaching. His confidence in using statistical methods can be gaged by a typical statement he wrote in "Educational Diagnosis" (1913):

> Tables of correlations seem dull, dry, unimpressive . . . . but only to those who miss their meaning. In the end they will contribute tenfold more to man's mastery of himself.⁴⁸

E.L. Thorndike and the Legacy of Teachers College

Teachers College became a hot bed of reform for education. Thorndike brought under his direction many psychologists who later came to emphasize the statistical approach in conducting research. Rudolf Pintner, one of Wundt’s graduates and later a professor at Teachers College, commented that: "No one man has had more to do with stimulating the measurement movement in this country than Professor Thorndike."⁴⁹

Thorndike petitioned for his appointment at Teachers College. He suggested that he would do two jobs for a salary just slightly higher than that of a "professor of psychology". He went on to argue,
If I am able to do a full years research work as a fellow and still attend thoroughly to the teaching work, you get as good a fellow and as good a teacher without paying for both. Whomever you get as an instructor in psychology will, if he is a fit man, want to spend all his spare time in research of his own. If your instructor is at the same time fellow, he will not do research of his own, but research under the auspices of and for the benefit of the Teachers College.

He was a practical man with very practical ambitions. He received the appointment and began a career where he applied research to the very practical problems that faced teachers and superintendents.

Thorndike's courses emphasized methodology. In proposing (in 1901) to offer a course to graduate students, he placed a familiar argument before Dean Russell:

It will be a two point course, would meet twice a week, and would be entitled "Application of psychological and statistical methods to educational theory and practice". The aim of the course would be to provide not only intending investigators in education with the means of making their studies effective, but also intending superintendents, principals, and other practical workers in education with means of combining with their administrative duties such methods of record keeping, physical and mental measurements, and teachers records, as would make all these of some real value to the individual
whose records were kept, to the teachers who kept them, to students in education who might later on go over the records in search for the answers to some educational problems. 51

He was concerned not only to instruct graduate students but also teachers and superintendents. In this way he could establish a particular form of methodological preference for data collection in the schools. The course would also ensure that school data would be recorded in such a way that it would be available to the types of methodological analyses that were being promoted by Thorndike.

At the same time Thorndike was urging Russell to accept his proposal for the graduate course, he was pressing for schools to allow him to test their students. For instance in 1901 he wrote to Professor Dutton, Superintendent of Teachers College Schools, requesting that he be allowed to apply a series of test batteries to students:

During the last three years I have been studying carefully the question of the utility of systematic measurements in connection with schools and am now able to offer to the administration of the Horace Mann School a definite plan superior by far to any advocated by psychologists. I have examined the different tests in use and tried others and have succeeded in eliminating those which have no significance as symptoms of general conditions. . .

If you think that it is wise to keep a record of the general growth toward maturity of children and of their standing in certain matters in comparison with
children in general of their age, I am willing from next year on to undertake the administration, recording and preservation of these tests and to send you annually a formal statement concerning the exceptional children found.\textsuperscript{52}

I don’t think it was incidental that Thorndike’s plan to enter the schools and his plan to offer a graduate course on testing and the statistical handling of data was coincidental. Thorndike’s students, their Ph.D. research projects, certainly drew on his ability to gain access to the schools. In October of 1902 Thorndike addressed a memo to Dean Russell proposing a plan to keep systematic records of the "mental condition and mental growth of the boys and girls in the Horace Mann and Speyer Schools". In this memo he also suggested that a "free clinic" be established to offer "educational advice to teachers and parents about exceptional children."\textsuperscript{53}

Thorndike provided both the courses and the applied context for his students. His courses emphasized the value of statistical compilation as well as the importance of test construction for mental assessment. Test items were useful if they discriminated between groups; the discrimination was based on simple statistical correlation. If a homogeneous group of students all get the item, the item stays. If the item correlation is low, it was rejected. In this way mental tests were constructed and standardized with large samples from New York schools.

Statistical norms were regarded as social norms. Research bolstered by statistical summaries, provided a
lever for school reform. It is interesting that many of the statistics textbooks written for psychologists and educators during the late teens and early twenties were almost exclusively written by Teachers College graduates or faculty.

In a short period of time Teachers College students dominated administrative positions in the schools. David Tyack and Elizabeth Hansot (1982) noted that during the period between 1900-1920s an educational trust was established to assert a meritocratic ideology into the school system. In 1915, they noted, a group of leading educators were brought together to assess the Cleveland Public Schools. Of the nineteen gathered together, eight were graduates of Teachers College and three others were on faculty there. They enjoyed each others discussion to such an extent that Charles Judd, one of the chosen, convinced the others that they should continue to meet-they became known as the Cleveland Conference. These few, which grew in number over the years, became the "political bosses" of the educational trust, the providers of placement for their underlings. There can be little doubt that Teachers College was influential in establishing the methods through which educational problems would be approached—both administratively and in terms of research perspective. Tyack and Hansot referred, quite appropriately, to Teachers College as the "West Point of the educational trust".

Educational research as it was practiced at Teachers College established a pattern for other departments of education. Certainly the most prominent University to
follow their example was Stanford. Ellwood Cubberley, a graduate of Teachers College, was invited to direct the Department of Education. He in turn invited Lewis M. Terman, in 1910, to join the faculty. Terman’s interests were strictly in the development of mental testing. In 1920 Terman recommended another Teachers College graduate to be added to their faculty--Truman Lee Kelley. Kelley was a promising statistician, he worked with Thorndike, was a professor at Teachers College, and was keenly interested in test development. Stanford was for all intents and purposes the Teachers College of the west coast.

The similarity between Teachers College and Stanford’s Department of Education was not so much one of organization, though this existed. Their outlook on educational problems and the methods employed in their research followed the example of their eastern alma mater. That is, their goal was to collect and compile as much data as they could. The aim was to test children in as many schools as was physically possible and to keep records of all the correlations. Through their research they would attempt to develop better mental tests and maintain a rigorous, thorough, program of standardization. It was necessary to their program to promote their methods so that more schools would allow researchers to enter their classrooms. They endeavoured to secure funds to set up an educational clinic so that test development could proceed on a population of exceptional children as well as offer a service to the public.

Although Lewis Terman was not a graduate of Teachers
College he was a sympathizer with their approach. He wrote in his autobiography:

One thing that I much needed and that Clark did not have to offer was instruction in statistical methods. It would have been an untold boon to me if I could have had a year with Thorndike immediately upon leaving Clark; but there were no post-doctorate fellowships in those days.  

Terman was convinced that Thorndike’s methods (or methods akin to them) were the correct ones to follow. He reasoned that for the present it was important to make practical gains at the expense of theoretical development. During the heyday of mental testing, I believe Terman accepted wholeheartedly the naive theory of measurement implied by the testing method.

Terman rejected the laboratory for the stated reason that he could not deal well with experimental apparatus. He found Thorndike’s work on testing to be stimulating, even though they differed on interpretation. That a paper & pencil test would reveal important data on the mental development of children was not questioned. That meaningful information about the child’s intelligence could be inferred from a test score, coupled with a statement of the probable error, was seen as unproblematic. He believed, as did many in the testing movement, that to be fully aware of the basis of such inferences, a thorough understanding of statistics was required. As far as Terman was concerned, psychologists needed training that would enable them to develop better tests and a training that would include a thorough education.
in the use of statistical methodology. Without statistics it would be difficult to assess the meaning of the large quantities of data that were collected within the schools.

The collection of data was a cooperative project. It required many individuals, not all of whom were trained psychologists. As can be seen in Thorndike’s program, he used graduate students and teachers to collect data. Terman established a similar program at Stanford. Records had to be kept by teachers and principals so their cooperation with the research objectives was necessary. As T.L. Kelley, Thorndike’s graduate student, wrote in the published version of his Ph.D. thesis:

The task of giving tests, establishing averages and calculating relations, which shall serve as the basis for prognosis of mental ability, is, in every sense, a social undertaking, and it is only because of the kindly cooperation of principals, teachers, and pupils of the two schools studied that it has been possible to secure the data that supply the material for this investigation.\(^61\)

As with all forms of cooperation, there is a social ordering. For the mental testing research in general this social ordering was based on a corporation model.

As Callahan (1962) has pointed out, the impact of business upon education was to assert a top-down management scheme. The factory metaphor was being applied to schools; business and industrial values of efficiency were used to assess curriculum effectiveness. Thorndike accepted these terms of assessment and devised methods to better ascertain
the effectiveness of teaching programs. Teachers College, and those who accepted their teachings, devised systematic ways of evaluating entire school divisions. One such program was the adoption of a year-by-year progress report which used percentages to evaluate gains in terms of efficiency in school programs. In the adoption of a top-down management scheme, Teachers College, and later Stanford University, exerted considerable influence on the development of curriculum in the education system.

In a top-down management program, there have to be those who are managed and those who report to management. In the system promoted by the Colleges, the professors were at the top. They directed the activities of their students and the good ones would be placed in good positions. They directed research through their graduate students. Often this research was carried out in schools where the superintendent was one of their graduates or a sympathizer with their aims. Also, many of these college professors sat on committees that directed funds to the schools or, more often, to the research projects that would have an effect on schools. They wrote reports, or, at the very least, had access to those who were writing the reports on the efficiency of a school division. Although this management system did not entirely replace the old network of political bossism that at one time had a say in who was elected to school boards and superintendencies, it had great appeal to local governments as it promised scientifically gaged reforms.62

The Managers of Method
An outstanding feature of the new movements in educational research was the manner in which they collected and generated data. Thorndike was a model for many others who would get involved in applied research, educational and otherwise. Thorndike devised graduate courses, gained entrance into classrooms, and had graduate students collect and analyse the data. He would teach the methods, they would employ them. He would then interpret the data or help the students to interpret the data. Cubberley, who had taken classes from Thorndike and George Strayer (who taught educational administration), followed a similar strategy for research.

Lewis Terman and Truman Kelley carried out their research programs on Thorndike’s model. They used their graduate students effectively to collect data and employ methods. Thus graduate classes at Stanford emphasized data collection and analysis. In their books, as well as any of the books their students wrote later, they never neglected to mention the cooperation of teachers, principals, and school pupils. They also stated that much of the data was collected by students in their courses. For example, Terman wrote in the preface to his book *The Measurement of Intelligence* that:

The Stanford revision of the scale is the result of a number of investigations, made possible by the cooperation of the author’s graduate students.\(^{63}\)

G.M. Ruch, a student at Stanford, wrote in his book *The Improvement of the Written Examination* (1924), that it was "an outgrowth of a syllabus prepared for use with a
university class in objective methods" and that acknowledgement of a number of the "author's students have been made throughout the text."64

The implication I want to draw from this is that the methods used by the graduate students were ones that they could master in a course. A long tenure of apprenticeship was not required to gain mastery over these methods. Thus, it was in both the instruction in methodology and in the interpretation of results that the professor maintained control over the research. The psychologist managed the entire knowledge generating activity. In emphasizing the management of research, I am suggesting that psychologists were not always directly involved in the collection of data nor were they in contact with the subjects which they studied. They did not observe, as a matter of course, the test taking activities of their subjects. They lacked the hands-on experience of doing research.

The psychologist taught methodology, engaged in public relations to gain access to new areas for research, promoted the objectivity of their new methods, and interpreted the data once it was gathered. The more tasks that could be assigned to assistants in both data gathering and data analysis (doing the statistical calculations) the more time these psychologists could spend in gaining funding, writing books, and doing committee work.

This pattern is evident in the work of psychologists at Teachers College and Stanford. Other universities, usually State Universities, followed suit. Certainly the pattern is clear in the works of Thorndike, Terman, and Kelley, all
leaders in the test development movement. Their efforts were concentrated on the writing of monographs and books.

Thorndike and Cubberley were productive textbook writers. The entire process of research and monograph production had economic renumeration for both the professors and the schools. E.L. Thorndike received a substantial earning from the sale of textbooks and tests. His biographer, Geraldine Joncich (1968), noted that although his salary was $10,000 in 1921, his total income for the year was $23,000. Most of the additional $13,000 was from the sale of books and tests. In 1922 his earnings increased to $39,000 and in 1924 his income was $68,000, five times his academic salary. Ellwood Cubberley earned extra money as an editor and writer of textbooks. He edited 103 of the 110 books in Houghton Mifflin’s popular education series. Tyack and Hansot (1982) commented:

By finding young scholars and commissioning their work, he also helped to create scholarly reputations and to establish his position as elder statesman of the field of education. Through the series [Houghton Mifflin] he gained power to anoint the new and make it respectable, to define the new science of education. And in the process he built up a fortune which he increased by wise investment in the stock market.

He used his wealth to fund a new education building and other projects which would benefit Stanford.

Of the many books Terman authored, the ones dealing with the Stanford Binet test and his long term studies on gifted
children (which generated several volumes as well as other books which were offshoots of this study, e.g. Terman & Lima Children's Reading) are good examples of the pattern. Kelley produced several textbooks on statistical methods and test interpretation throughout the twenties.

Statistical methods were of fundamental importance to these research programs. They had to deal with and derive meaning from large amounts of data. Their theory of measurement was simple and this enabled quantitative values to take on a straightforward meaning when they were treated statistically. Statistics were interpreted as tools that led to an uncomplicated disclosure of what rested behind the visible data. Although statistical concepts in themselves were difficult to master, once mastered their application to mental test data proceeded unhindered. Statistics became not only statements about measurements but, in an important sense, became measurements in themselves. The psychologist-statisticians and those they served claimed that statistical methods could resolve all the differences between the various mental tests and had the power to arbitrate on decisions. In 1924 Lewis Terman wrote:

The statistical treatment of mental test data has answered once for all the question regarding their validity. Time and again it has been shown that the scores on an intelligence examination enable us to predict college success as accurately as we can predict it from four years of high school marks. He suggested further that the "value of the prediction may be judged wholly in terms of correlation coefficients or
other quantitative, objective evidence." In 1927, in the introduction to Kelley’s textbook, *Interpretation of Educational Measurement*, Lewis Terman commented that: "When we become as conscious of the probable error as Professor Kelley would have us, our tests are certain to undergo rapid and marked improvements." Kelley, in the same text-writing on the history of mental measurement – urged that the validity of educational tests had fallen behind studies of reliability because of the difficulty of arriving at a statistical evaluation of validity problems. He suggested that school boards at times have used tests that, for one reason or another, appealed to them and have not allowed their validity to be challenged. He continued:

If the deliberations of such boards can be supplemented by an adequate statistical technique, the problem of the validity of a test will shortly assume the importance that is its due. Statistical method, it was thought, could bring harmony and unity of purpose to the field of mental testing as well as to psychology.

Statistical methodology allowed psychologists to carry out investigations in the "real world" and justified the testing of groups. Rudolf Pintner noted that as tests for different mental processes multiplied, the group method became popular. He also suggested that Thorndike "was among the first to see the advantages of this method" and should be considered "the leader in the movement." Pintner saw the connection between group testing and statistical methodology. He commented: "The use of this method was
greatly stimulated by the investigations of the educational psychologist in the school room, and also by the growing interest in the relationship between different mental processes studied by the mathematical formulae for correlation." The laboratory was not the only place to derive psychological knowledge. Not only did textbooks on statistics begin to appear in greater number, as did courses in statistics, but journals devoted more space to discussions of statistical methods.

The model of the expert as a ‘manager of methods’ asserted itself strongest among the mental testers. In many respects this makes sense. They followed a management model in their research. They organized the work and set the problems (as well as the data analysis) for their graduate student assistants. They also spent a good portion of their time doing public relations work to promote their methods. They served on many committees, most of which were concerned with reform issues or funding agencies. They perceived their duty as that of a scientist administrator.

As managers of methods it was required that their lack of close contact with the subjects who participated in their studies would not be detrimental to their interpretation of the results. Thus, methods were attributed properties that enabled them to carry a good supply of information. Perhaps the most important of these methods was statistics. In the chapters which follow, I attempt to demonstrate how statistics were interpreted by mental testers to take on inferential capacities and the debates that emerged as a result of these interpretations.
Conclusion

The focus of this chapter has been on the two models of expert which were outlined earlier. I attempted to place these models within the contexts in which they operated. There is of course considerable overlap in the two models. Clearly the "objective arbitrator" experts at times called upon method to assert the unbiased nature of the knowledge. Also the "manger of methods" experts would speak of their judgements concerning the data. I think, however, that each group conceived of methods in ways that promoted their own special interests.

For the "arbitrator" experts, method was used to isolate and describe the "facts". The entire context was taken into account. They were as interested in the effects the method had on a subject as the data generated through it. To learn the craft knowledge of science was a long process and such knowledge was learned as much through participation in experimenting as through specific instruction. As Boring commented, one had to learn to breathe the atmosphere of the laboratory, absorb the culture, work the eighty hour week.

The close environment of the laboratory encouraged sophisticated discussions of apparatus and methods. These discussions appealed to a limited audience and there was no urgency to promote these methods outside of the laboratory. What they did promote was the notion that a good scientist produced factual knowledge. If one could find applications for the laboratory derived facts, all the better. Many psychologists endeavoured to do just that and, for the most
part, they were not discouraged by the community of laboratory psychologists. What was objectionable to the laboratory researcher was using a standpoint based on practicality to determine the course of laboratory investigations. The focus for the scientist, it was believed, rested with the pursuit of psychological principles for their own sake.

The 'managers of methods' approach placed the research activity outside of the laboratory. Here the psychologist faced a critical audience different from the one faced by the laboratory psychologist. Teachers, principals, school boards, industrialists, all could examine the efficacy of their methods. These outsiders criticized their methods from a standpoint that was not shared by the psychologist. Sometimes they performed these critiques with considerable effect, as in the case of Walter Lippmann.70 Thus, efforts were made to address critics and to endeavour to popularize the methods used. Terman ran into difficulties addressing the criticisms of Lippmann, but he regarded them as strategical. He vowed to never again enter a debate in the public forum. He wrote to Jessie Fenton:

As for myself, I think that answers in the future will be confined to presentations of data in scientific periodicals. There is no use trying to argue with some people.71

Since these psychologists dealt directly with other professionals, it was not good enough to assert the unbiasedness of their own judgements. Rather, they adopted the strategy of placing in the methods themselves an
inferential capacity. Therefore, they portrayed themselves as skilled technicians.

This resulted in two problems. First, it raised the ire of their laboratory peers. Second, it suggested that if anyone uses the methods, they can act in place of the psychologist. Both had an effect on the development of mental testing methods.
Notes


4. C.E. Seashore, "Editorial: The Educational Efficiency Engineer". Educational Psychology (1913) 4, p. 244.


7. In 1920 E.G. Boring carried out a survey showing that the majority of psychologists were involved in experimental research, but it was a near majority. Lewis Terman responded to Boring’s survey by pointing out that if the data was tallied in a slightly different manner, it would indicate that the majority of psychologists were involved in applied research at some level. See Boring, "Statistics of the American Psychological Association", Psychological Bulletin (1920) 17, 271-278; & Terman, "The Status of Applied Psychology in the United States, Journal of Applied Psychology, (1921) 5 1-4.


9. Kuhn used the term "exemplar" in the postscript to The Structure of Scientific Revolutions (second edition), Chicago: University of Chicago Press, 1970, p. 187. He wrote: "By it [exemplar] I mean, initially, the concrete problem-solutions that students encounter from the start of their scientific education, whether in laboratories, on examinations, or at the ends of chapters in science texts. To these shared examples should, however, be added at least some of the technical problem-solutions found in the periodical literature that scientists encounter during their post-education research careers and that also show them by example how their job is to be done." (p. 187)

For Kuhn’s discussion of the role of text books in science see, "The Function of Measurement in Modern Physical Science", in The Essential Tension, Chicago: University of
10. Nothing was further from the truth. Although there was agreement as to tables of content, the putative 'facts' that were presented by different authors were quite different. Psychologists disagreed with one another about methods as well as content. Yerkes conducted a survey of scientists to see if they would rate psychology as a science; their verdict was negative. He concluded:

The sad truth is that today psychology means very different things even to psychologists themselves. . . As a group we lack the strength of faith in our aims, methods, and ability which alone makes for success in research. We lack enthusiasm; we are divided; we waver in our aims; we mistrust our methods as well as our assumptions; we question the value of every step forward, and, as an inevitable result, our subject lags at the very threshold to the kingdom of the sciences. (R.M. Yerkes, "Psychology in its relations to biology", Journal of Philosophy, Psychology and Scientific Method (1910) 7 p. 122.)


does some violence to the traditional classification of psychological topics and to their conventional treatment. For example, the reader will find no discussion of consciousness and no reference to such terms as sensation, perception, attention, will, image and the like. These terms are in good repute, but I have found that I can get along without them both in carrying out investigations and in presenting psychology as a system to my students." (p. viii).


17. Loren Baritz, Servants of Power, op.cit., remarked: "scientific management not only conditioned the industrial climate for the psychologists, it determined to a large degree the direction, scope, and nature of psychological research." (p. 35) David Noble (America By Design, op.
cit.) outlines the influence of scientific management and suggests some of the effects it had on psychology; see his chapter on "A Technology of Social Production", pp 257-324.


19. Kurt Danzinger in "Educational Administration and a Critical Shift in Psychological Research Practice" (presented at Cheiron Meeting, Vassar: 1984.) argued that statistics were useful as a tool in condensing large amounts of data from the classroom and thus statistics were revealed as a useful methodology for making administrative decisions. The pattern of the use of statistics by psychologists was apparent in the early work of Thorndike and Cubberley who were involved in the administrative reforms that appealed to notions of scientific management. For a fuller discussion see D. Tyack and E. Hansot, Managers of Virtue: Public School Leadership in America, 1820-1980. New York: Basic Books, 1982.


22. Although the specific point that behaviorism functioned as a more efficient route to applications has not been clearly made in the literature, O'Donnel has convincingly shown in his doctoral dissertation that the rise of
behaviorism was an important development in applied research. He noted that Watson attempted to unite the discipline around a definition of psychology that placed the prediction and control of behaviour at its center. O’Donnel wrote: "Behaviorism purported to be the science that investigated outcomes - defined as behavioral responses. Therefore, behaviorism did not merely assist applied psychology, as Watson had argued, it was applied psychology." (p. 540). See O’Donnel, The Origins of Behaviorism: American Psychology, 1870-1920, op. cit.. Also Lucille C. Birnbaum in "Behaviorism in the 1920s" (American Quarterly, 7 1955, 15-30) stated that Munsterberg endorsed behaviorism as a promising vehicle for applied psychology.


Development., in preparation (1985). Danzinger noted that psychologist’s "most impressive successes" in applying their services was in the educational system. He commented: "The education industry not only provided psychologists with jobs and a recognized social function, it also came to exert a major influence on psychological research, both in terms of subject matter and in terms of methodology." (p. 11)


27. I have drawn upon Lawrence Veysey’s excellant work, The Emergence of the American University, (Chicago: University of Chicago Press, 1965; pp 180-259) for the discussion which follows. I have tried to tie in certain aspects of psychology’s development to some of the points raised by Veysey.

28. Andrew F. West referred to the growing tendency to examine the quality of academics by their research record was a "destructive theory". See his article "The Changing Conception of 'The Faculty' in American Universities." Educational Review (1906) 32 p. 11. Irving Babbitt suggested that the Ph.D. degree led to "loss of mental balance". See Literature and the American College. Boston:
29. Edwin Slosson, *Great American Universities*. New York: Macmillan Company, 1910. pp 509-516. He talks about the "warfare between science and classics" as "practically over". (p. 509). Classics was used in a broad sense and included the new studies of Literature, thus I referred to 'humanities' rather than 'classics'.


36. J. Royce, *Outlines of Psychology*. op. cit., p. 18, the emphasis was in text.


38. For instance, James McKeen Cattell remarked: "It cannot be denied that the organization of the departments of a university is one of the difficult problems that confront us. The German plan, according to which the individual rather than the department is the unit, is in many ways preferable." See *University Control* (New York: 1913) excerpts reprinted in Richard Hofstadter and Wilson Smith (editors) *American Higher Education: A Documentary History*. (Vol. II) Chicago: University Chicago Press, 1961, 784-808, p. 800. E.B. Titchener also upheld the ideal of the German University, see E.G. Boring’s obituary of Titchener, *American Journal of Psychology*. (1927), 28 489-506.


42. Ibid., p. 22.


45. E.L. Thorndike, "Measurement in Education", Teacher's College Record. (1921) 22 371-379, p. 371. For President Russell's comment, see a manuscript sent to Thorndike written by Russell and dated August 22, 1949, Archives at Teacher's College, New York.


52. Thorndike to Professor Dutton, November 27, 1901. Teacher's College, Special Collections, Thorndike Papers. He listed many tests, for instance, "tests of control over the muscles, rate of writing, rate of vibratory movement, rate of precise movement guided by the eyes, accuracy of full arm movement, hand balance--these made up the motor movement tests. He had tests of "perception, association,
attention and memory" and "tests of practical judgement". Of these latter tests he said, "These measures would give, as nearly as any brief examination of an individual could do, an idea of a pupil's standing in ingenuity, originality and active reasoning powers."

53. E.L. Thorndike to James Russell, October 7, 1902. Teacher's College Library, Special Collections, E.L. Thorndike Papers.


56. Ibid., p. 131.

57. Ibid., p. 126.


59. In his autobiography, Terman commented that although his contributions to the theory of mental measurement were "not great", the significance and value of his work rested with his development of mental test. "I think that I early
saw more clearly than others", he wrote, "the possibilities of mentality testing, have succeeded in devising tests that work better than their competitors, . . ." (p. 328, op. cit.)

60. Ibid., p. 311, footnote 4.


62. Both Raymond Callahan, *Education and the Cult of Efficiency*, and Tyack & Hansot, *The Managers of Virtue*, have given detailed accounts of the business of education and the importance of the "educational trust" system which came to dominate American Educational organization during the first three decades of this century. Tyack & Hansot have shown the importance of the placement system of graduates and how a few professors managed to exercise considerable influence on the development of educational values.


65. Geraldine Joncich, *The Sane Positivist: A Biography of E.L. Thorndike*. Middletown: Wesleyan University Press, 1968, p. 399-400. It is worth pointing out that much of the increase in his earnings was due to his success as a
textbook writer. As Joncich pointed out, the adoption of his text *Arithmetics* (1917) was adopted into the Indiana State education program, which alone would have added $15,000 in royalties to his income.


70. Walter Lippmann’s articles are well known. He attacked the IQ test and the testers in a series of articles in the *New Republic*: "The Mental Age of Americans" (Oct. 25, 1922); "The Mystery of the ‘A’ Man" (Nov. 1, 1922); "The Reliability of Intelligence Tests" (Nov. 8, 1922); "The Abuse of the Tests" (Nov. 15, 1922); "Tests of Hereditary Intelligence" (Nov. 22, 1922); and "A Future for the Tests" (Nov. 29, 1922). All of these have been reprinted in Clarence J. Karier, *Shaping the American Educational State*, New York: The Free Press, 1975, pp. 283-305. Terman responded to these criticisms. See "The Great Conspiracy or the Impulse Imperious of Intelligence Testers, Psychoanalyzed and Exposed by Mr. Lippmann," *New Republic*, 111
Dec. 27, 1922, 116-120. Terman found the exchange very unpleasant and vowed to never again to enter into public debate.

71. Terman to J.C. Fenton, March 12, 1923. Terman Papers, Stanford University Archives.
The Psychological Bulletin began in 1912 to run a series discussing new uses for the coefficient for correlation. James Burt Miner was the commentator. In the first article of the series Miner drew attention to a monograph by W. Betz, Über Korrelation (1911). He pointed out that Betz argued that "correlation alone does not demonstrate a functional connection. . . . An inventory of correlations cannot disclose psychological secrets unless supplemented by an understanding of mental facts."¹ Correlations did not, in other words, allow a researcher to make causal inferences. Yet, just three year later, Miner wrote in the same series that: "When we examine the improvements which have been made since the last resume, we find ourselves getting notably nearer to the goal of measuring causal relations through correlation."²

More than anything else Miner was reflecting the enthusiasm for the new technique. Psychologists were close to viewing the correlation coefficient in causal terms. E.L. Thorndike in his text An Introduction to the Theory of Mental and Social Measurement reasoned that correlation suggested that cause "must be equal to effect." He went on to say: "Whenever one finds two qualities correlated he may properly proceed to test the hypothesis that one caused the other in part and that both are due to some common cause."³ A.S. Otis, a student of Terman’s, wrote that a "coefficient of correlation between two series of values is a measure of
the percentage of elemental causes common to both."4 In his review of correlational research for 1918, Miner commented on Harold Rugg's new text on statistics:

The methods of computing the various coefficients are also presented simply and fully. Unfortunately, space forbade his including partial correlation which illuminates the whole subject of analysis of causes. It will probably be the most important feature of the work in correlation during the next decade.5

But not all psychologists were enthusiastic about the use of correlation statistics, nor were other professionals who made use of statistical summaries.

Truman Kelley had developed a "machine" for calculating correlations. He wanted to gain financial support and make the machine available to other professionals using statistics. What is revealing in his attempt to gain funds was that other professionals were not interested in supporting his calculating machine. He petitioned Lt. Col. Leonard P. Ayers (Office of the Chief of Staff, Statistics Branch, in the War Department) for support. Kelley wrote that it had been his "intention to patent this machine" and to get it out himself. He was willing, however, to "forego any personal rights in the machine" if he could find "a way to make it quickly available". He reasoned that it was important to get the machine out quickly since there was a great demand for "correlation work in psychological and trade test work in the army, as well as probably other army fields." He wrote Ayers asking if their department could
finance it, or if this was not feasible, whether he could provide a statement as to the government agencies which "have need for correlation work". Kelley proposed that with this information he could go to the Trustees of the Carnegie Corporation and request financial assistance. 6

Ayers response must have been disappointing to Kelley. Ayers wrote a lengthy letter to Kelley saying that he surveyed many involved in the Statistics Branch. He wrote:

I much regret . . . that I do not believe that my branch of the work here or any related Army statistical office would have a sufficient amount of work to do involving correlations to justify it in financing the production of a correlation machine.

He went on to offer the details of his survey:

Our own division has been in existence for more than a year and has had a membership of 60 or 70 people during a large part of that time. I find, on investigation, that there is no record of any correlations having been figured by anyone in the force during that entire period.

Not being satisfied with offering just his experiences, he polled Raymond Pearl, who was the statistician for the Food Administration, for his comments. Ayers wrote that even though Pearl’s department was "now several hundred in number", he [Pearl] reports that "since the beginning of their work a year ago they have figured one correlation and that he does not think they will have more than one or two a year." Ayers also remarked that the "statistics offices in the Ordnance and Quartermaster Departments, which have
forces of several hundreds of workers, have never had occasion to compute a correlation." Ayers also stated in his letter that all of these offices had statisticians who were "entirely familiar" with the technique of correlation; they simply did not use the statistic very often in their work.

A growing number of psychologists, however, were committed to the calculation of correlations. They began to devise short cut methods of calculation. In 1917 L.L. Thurstone proposed a method of calculation that was easily adapted to an adding machine as it did not require the direct calculation of deviations. T.L. Kelley (1916), at the request of Thorndike, developed tables to facilitate the calculation of partial correlations and regression equations. In the introduction to this work, Thorndike figured that Kelley’s tables would reduce the labor of calculation by 80 percent. Others such as H.F. Adams (1917), J.C. Chapman (1919), L.P. Ayres (1920) and H.A. Toops (1921) all set out short methods for calculating coefficients of correlation. H.E. Burtt (1919) showed how correlations could be calculated using a slide ruler and Toops (1921) produced a "correlation plotting device" and the requisite printed sheets for the "device". T.L. Kelley produced a "correlational-biometricians slide ruler" in 1920 and approached the Keuffel-Esser Company (specialist in the production of slide rulers, drawing materials, mathematical and surveying instruments, measuring tapes, etc.) to manufacture it. After consideration the president of the company wrote to Kelley saying that "technical
difficulties and the cost of placing [it] on the market" were "so great that we do not consider the proposition a favorable one from the commercial standpoint."\textsuperscript{12}

Psychologists were almost alone in promoting correlational methods in America. Their development of short cut methods in calculation were often derived from the Biometricians under Pearson's direction. That they seldom credited these statisticians was noted by Beardsley Ruml. He criticized psychologist-statisticians for their unwillingness to cite their sources and blamed the editors of some of the new educational journals for allowing such unscholarly activities to occur. Ruml (1921) wrote:

The authors of such papers themselves can not be held blameless; but after all it is the journals that have brought into existence the reputations to which school superintendents turn for advice.\textsuperscript{13}

Psychologists were interested in developing quicker, more efficient, means of calculation. They were also keen to make reputations. This concern for recognition, as one might expect, led to a number of priority disputes between psychologist-statisticians over the "discovery" of certain formulae. There were, of course, clear renumerative benefits that accompanied both a strong reputation and the wide acceptance of any calculating apparatus.

Although mental testing and its dependence on statistics was widely accepted as legitimate methodology, there were detractors. These critics were leery of the scientific status of mental testing and were severe in their assessment of the manner in which statistical methods were being used.
It was these criticisms of statistics that were clearly the most damaging. Such critiques challenged the scientific status of the methods, and by implication the methodologists, of mental assessment.

What rested behind their criticisms was a rejection of mental tester’s knowledge generating practices. An industry of testing was emerging and it justified its scientific status largely through its interpretation of statistical methodology. The more psychologists explored the uses of statistics, the more they read into them a capacity for drawing out scientific inferences.

The mental testers were opening new vistas for psychological science and were not to be deterred. Lewis Terman and Truman Lee Kelley, for the most part, defended the scientific status of mental testing and its associated methods. They answered the critics, sometimes with elaborate articles. At other times they set out arguments that promoted their methods as another branch of scientific psychology.

In this chapter I examine the critics of mental testing methods. As it was the use tester’s made of statistics that drew the most sophisticated criticism, as well as the most devastating, I concentrate on those critics who set out these arguments. Also, it is easy to see in these criticisms - the brunt of which was to remove the inferential power ascribed to statistics - was a rejection of a particular research practice. Had the success of these critics been overwhelming, the implications for the ways in which psychology would proceed to apply its knowledge (and
the way in which it would conceive of the "psychologist-expert") would have been vast.

The Critics:

There were many who criticized the mental testing movement. Pintner recalled that many non-psychologists were opposed to mental testing and among psychologists there were some who regarded it as a fad that would disappear. He remarked:

On the whole . . . the controversy that raged around the Binet Scale left us with clearer ideas as to the problems and methods of intelligence testing, even although much of it was mere opinion and useless. He suggested that it "might be amusing" to enter into a detailed account of the criticisms, but most of it "has no application at the present time. Most of the issues, he remarked, had "already faded into the historic past" at a rate "so rapidly as to make one smile at the emotional outbursts and personal feelings that were aroused."14 But many of the critics were not amused, nor had their critiques faded into the historic past. Many of these critics, who were psychologists, questioned the manner in which statistics were being used by the mental testers.

Two principal critics are examined here, Edwin G. Boring and Beardsley Ruml. The reason for selecting these two rests with three considerations. First, both wrote a series of articles remonstrating the mental testers by attacking their use of statistical methods. Second, both received rebuttals from Lewis Terman and Truman Kelley (who were
leaders in the testing field). Third, both Ruml and Boring held positions of influence and therefore carried some political clout.

Beardsley Ruml

Ruml received his Ph.D. at the University of Chicago. His thesis examined the reliability of mental tests. After completing his thesis in 1917 he became Walter Bingham’s assistant at the Carnegie Institute of Technology. Here he worked in the area of applied psychology and became acquainted with both Walter Dill Scott and L.L. Thurstone, who was a young instructor there at this time. During the First World War he served on the Committee on Classification of Personnel in the Adjutant-Generals office. Scott was chairman of this committee and it is likely that Ruml met Kelley here since they served on the same committee. After the war, Ruml moved to Philadelphia where he joined W.D. Scott in the formation of the Scott Company. He left the Company in 1920, as did Scott, and became assistant to James Angell (his Ph.D. supervisor) who was now president of the Carnegie Corporation. In 1921 Angell accepted the presidency of Yale University and Ruml began to work for John D. Rockefeller Jr.. Later in that same year, due to the persistent recommendations by Angell and Abraham Flexner, Ruml was recommended to JDR as Director of the Laura Spelman Rockefeller Memorial Fund. In May, 1922, Ruml became--at the age of 27--director of the Spelman Memorial and remained as such until the Fund was incorporated into the Rockefeller Foundation in 1929.
After the Spelman Memorial no longer functioned as an independent body, Ruml moved to Washington and served as an assistant to Arthur Woods on President Hoover's Federal Committee on Employment. In 1931 he returned to Chicago to become dean of social sciences. He remained there until 1934 when he departed for New York to become treasurer for the R.H. Macy Company. In 1945 he became chairman of the board.

I am interested in Ruml's early career as it was at this time when he criticized the methods employed by mental testers, particularly their statistical methods. In 1920 he wrote an article titled "The Need for an Examination of Certain Hypotheses in Mental Tests" and followed this with a 1921 article, "Reconstruction in Mental Tests". Both were critical of the use of statistics, but from slightly different perspectives.

In the first article his concern was that although mental tests had been of some practical value, they made no significant contribution to a scientific theory of intelligence. He believed that the development of a sound theory was necessary and was of paramount importance because of the wide use of mental tests. In this article he speculated that the reason for this lack of theoretical development rested with tester's inclination to accept as their data "derivative facts", rather than "manifestations of intelligence as we commonly experience them."

"Derivative facts" were the results of subjects' mental test performance and were "fundamentally biased", he thought, due to "the type of analysis which our limited and frequently misused statistical techniques" allowed.
Ruml argued that the mental testers accepted certain statistical assumptions as valid because they wanted to use statistical techniques, not because these were empirical attributes of the data. He commented:

When the results of several tests are combined, as for example, in the Binet series or the Army Intelligence tests, the standing in the combination is again expressed in terms of a linear scale, not because we have analyzed our concept of and experiences with general intelligence and have found it so expressible, but because our common methods of test measurement and combination preclude any other result.

Indeed, Ruml pointed out that his experience with trade tests indicated that the data departed significantly from linearity. He wrote to T.L. Kelley after publishing his paper:

I am afraid I am stating as gospel truth the prevalence of non-linearity in trade test question relationships, and no methods with which I am acquainted will produce proper rectification.

Ruml pointed out in his article that despite available data which indicated non-linearity, psychologists had a preference to use coefficients of correlation - "with the implied assumption of linearity" - on their mental test data. He complained that testers should draw out the implications of assuming linear regressions between test performance and general intelligence before proceeding further. He urged that since the trade tests found the use
of partial correlation methods to be "inapplicable" because of the lack of linearity, testers should be circumspect in using such methods in mental tests. "[W]e are building on the sand as long as the consequences of such an assumption [linearity] are not critically examined." This was Ruml's indictment of the mental testing field.21

In his second paper, "Reconstruction in Mental Tests" (1921), his approach was similar but his emphasis shifted slightly. This is due in part to the fact that the paper was written in response to S.L. Pressey's (1920) note "Suggestions Looking Toward a Fundamental Revision of Current Statistical Procedure, as Applied to Tests".22 Ruml was concerned that Pressey was "traveling over dangerous ground toward a goal that is hardly wholesome." He added that he believed that Pressey's point of view was shared by other psychologists who were involved in the testing field.23 Ruml felt that his concern over Pressey's note was more important than it might first appear.

Of central concern was Pressey's insistence that what was of primary importance was practical prediction. All test methods, according to Pressey, should be judged by their ability to predict in the context in which they were applied. Thus if a statistical method aided prediction it should be used. If it did not serve such practical considerations, it should be rejected. Thus Pressey believed that the time spent on determining, through partial correlational techniques, the statistical weightings to be assigned to different components of a test was a waste of time. Such statistical sophistication, he reasoned, was
useless unless there was a practical, predictive payout. Also, if a statistical technique proved useful for prediction it did not matter if the data violated important statistical assumptions. As Pressey wrote, "The actual distribution of various traits is a matter of academic interest only."  

Ruml was concerned that questions of the reliability and validity of methods were subordinate to questions of practicality. Even though he was critical of the use of statistical methods in mental measurement, his strictures did not call for an abandonment of such methods, but for a critical evaluation as to what they actually tell us about intelligence. Thus statistics should be judged according to whether or not they inform us as to the real underlying dimensions of mental activity. Of Pressey's standpoint he commented:

... if we make practical efficiency the criterion by which our statistical technique is judged, we shall encourage the use of innumerable methods, faulty or merely expedient, which have, to be sure, served a purpose, but which can not be genuinely productive in a scientific sense.

Ruml's second concern was that using a practical perspective as the sole criterion for determining the acceptability of a method "results naturally in a tendency to ignore the necessity for analysis, and for the isolation of variable factors." The result was, he thought, that there would be a growing emphasis on "omnibus tests" and methods which were "concerned only with total score."
applied psychological science was to develop, Ruml thought that applied methods would have to be more rigorous. He wrote:

The future of mental tests, even as applied science, hinges on the capacity of the field to produce contributions that will give us more light on the general problems of mental adjustment. These contributions will be in the highest sense of theoretical importance. It is therefore, to my mind, unfortunate that there should exist a point of view towards statistical technique that will ultimately bring about increased scientific sterility of the field.  

Both papers by Ruml presented arguments that were critical of the indiscriminate use of statistical methodology. The first suggested that one must critically assess the implications of violating statistical assumptions with respect to the meaning of the data. The second paper pointed out that it would be inappropriate to judge the efficacy of statistical methods from the perspective of their practical, predictive value. That some testers were doing just that was of deep concern to Ruml.

In part this concern stemmed from the fact that Ruml viewed statistical methodology as providing a useful means for arriving at scientifically valid conclusions, with the proviso that they are used appropriately. If they are used without caution as to what they imply about the data, they can mislead rather than lead the researcher. Both of these arguments fit with Ruml’s overall perspective of applied
science.

In an unpublished paper Ruml read before the New York Academy of Sciences, Section of Anthropology and Psychology, he set out a statement of his position on matters of applied psychology.\(^2\)\(^8\) Although he agreed that in applied research the choice of method rested with considerations of cost, he did not see this as settling for crude techniques. He argued that although the technique chosen may not be as accurate as another because of its economy, it may be of a sufficient calibre to answer the problem faced by an industry. He did not think that these economic considerations that the applied psychologist could settle for less than an accurate depiction of the data. Ruml commented:

The psychologist working in industry is vitally interested in accuracy of method; but his interest is in whether this accuracy costs more than it earns.

He went on to clarify his point:
The comparison of laboratory with practical procedures rests on something more fundamental and something quite different from the contrast between exact methods and rough, between the precise and the approximate. The elements of quantitative exactness, of accurate insight, of imagination, are present in both cases.

Where the mode of operation differed between laboratory and practical procedures was, according to Ruml, "the criterion of selection" of a particular method.\(^2\)\(^9\)
Ruml pointed out that it was wrong to think that laboratory methods stood for exactness while applied methods suggested looseness, lack of precision, or approximation. That the two approaches differed was clear, but Ruml insisted that:

. . . the methods of applied industrial psychology are, or at least certainly ought to be, chosen with as much discrimination as are the methods of the laboratory, after a study of all the factors influencing choice of method that are pertinent to laboratory investigation--and one factor in addition, the dollar return."

Ruml's concern with mental testing - and the use being made of statistics therein - represented a hesitation on his part to think that mental testers were being careful and scientific in the evaluation of their methodological procedures.

Not only was Ruml concerned about the scientific value of the work being done in the mental testing field, but he was equally concerned about the practical value of the work. He saw the mental testers as choosing methods that appeared to simplify prediction, but their endeavours were not less costly. They collected piles of data, yet the practical and scientific worth of it was questionable. He wrote:

The piling up of data has therefore been of little advantage, in fact it has created a wilderness of tangled issues of trifling importance removing still further the possibility of theoretical evaluation and interpretation.30
Ruml, like other critics of mental testing, felt that methods were being adopted for reasons other than their scientific efficacy or their practical validity. As far as Ruml was concerned, the mental testers marched on with their research despite their failure to provide anything of scientific value and little in way of practical value, given the degree of research activity. This progression disturbed Ruml.

Ruml was an applied psychologist, one might say even more of one than many of the mental testers. His concern was that research should be designed to address problems efficiently and effectively. In such a system, there was no room for favouring one methodology over another as researchers had to be eclectic and creative in their approach. Ruml was clearly perplexed by his mental testing colleagues as they used statistical methodology in all situations. As he wrote to Kelley, it seemed as if they wanted to use statistics and would inadvertently distort their data to do so.

Ruml’s criticisms of the use of statistics were not so much a rejection of the usefulness of these methods. Rather he rejected applied research that appeared to him to serve interests other than that of generating trustworthy knowledge through cost-effective methods. To Ruml it appeared that mental testers were using statistical methods not so much to establish a practical knowledge base as to advance and promote the use of statistics in applied research.

Applied psychologists such as Ruml, Bingham, and Scott
were concerned that their colleagues believed that applied science was less than rigorous in its approach. The thought that large sections of the mental testing movement lent credence to such a notion. Thus individuals like Ruml, though clearly applied psychologists, came out and openly criticized the methods of mental testers. Their concern was not distinguishing between pure and applied research but rather between applied research that was rigorous in its methods and applied research that was not. The right to generate knowledge, even if it be applied knowledge, had to be earned through a dedication to analytic research. There were no short-cuts.

After writing these criticisms, both of which were rebutted, Ruml became— in 1922—the director of the Laura Spelman Rockefeller Memorial. He endeavoured to shift the emphasis of this foundation from its original charitable and philanthropic purposes to a support system for socially directed scientific research. In October of 1922 he indicated in his "General Memorandum" just where he would direct the monies of the Memorial. He wrote:

Experience seems to show clearly that the results of investigations in the social sciences, where they are conducted by obviously impartial scientific agencies, and where these results are generally accepted by scientific men, come to play a definite and wholesome part in the thinking of people generally. It hardly seems too great an assumption to conclude that additions to the body of scientific knowledge in the social field will also have their
due influence on public welfare.\textsuperscript{31}

He believed that certain areas in the mental testing field did not proceed with the support of "scientific men". Research programs that did not appear to be wise in their use of mental test methods were unlikely candidates for support from the Laura Spelman Rockefeller Memorial.

Edwin G. Boring

It could be argued that Boring was the most thorough critic of mental testing during the 1920s. He was certainly a long lasting one. He wrote criticisms of the use of statistics in psychological research - in mental testing in particular - in 1919, 1920, 1926, 1941, 1960 and in 1961. Boring was regarded as an outstanding experimentalist and, of his generation, he became the most outstanding promoter of experimental psychology. When he first began to formally criticize mental testing he was Head of the experimental laboratory at Clark University, an institution modeled on the German Universities and dedicated to research and graduate education. In 1922 he left Clark, under unfavorable conditions of departure, and assumed charge of the Laboratory at Harvard University. He lived out the rest of his career at Harvard. Boring was active in the American Psychological Association becoming its President in 1928.

Like Ruml, Boring believed that mental tester’s reliance on statistical assumptions reduced the scientific merits of their results. He wrote to the biometrician Raymond Pearl in 1920:

Seriously our mental testers are wasting a glorious
amount of energy on figuring out results that don’t mean a damn thing when they are gotten.\textsuperscript{32}

Boring claimed that his interest in statistical inference was stimulated by his growing dismay over the methodological practices of mental testers. The claim he objected to was that they thought true scales of intelligence could be discovered by seeing how an arbitrary test scale could be "altered to make the distribution of a large homogeneous sample normal."\textsuperscript{33} It was through his association with the World War I testing program that Boring came to doubt the reasoning that gave priority to what he termed "the \textit{a priori} use of the Normal Law of Error as an hypothesis in scaling mental abilities and in building up statistical methods."\textsuperscript{34}

Boring recalled in his "Biographical Sketch" that it was psychologists’ willingness to participate in a program of research that appeared so wrong headed that eventually pushed him into the "Congressional Library" to read "Todhunter’s History of the Theory of Probability, and Laplace and Gauss, and then Venn and von Kries and some others on Wahrscheinlichkeitsrechnung."\textsuperscript{35}

As with most autobiographical commentary, it is never comprehensive. The details are sometimes deleted in favor of the unity of a perspective. Boring neglected to mention Raymond Pearl as an instigator of his investigations into the logic of statistical assumptions as they were being applied in psychological research. In 1920 he wrote to Pearl:

\textit{You may not recognize it but there is a certain}
amount of inspiration gathered from yourself in the one [paper] in Amer. J. Psychol. This paper is supposed to be a psychological rendering of the strictures that you made upon the statistical attempts of Yerkes' crowd in the Surgeon General's Office, of which I was one.36

Although Boring may have gained inspiration from Pearl, when he arrived to work with Yerkes on mental tests he brought along a scientific orientation that was from the outset critical of such research. He arrived fresh from his tutelage under E.B. Titchener at Cornell. Boring noted in his autobiography that: "Titchener's in-group at Cornell had appreciated mental-testers in much the same way that the Crusaders, gathered around Richard Coeur-de-lion, appreciated Moslems."37

In his autobiography he commented that he respected the dedication of the mental testers and noted that they "resemble the experimentalist in habits of work."38 To interpret this as support for mental testing research would be an error.39 He admired their work habits, not their work. As far as Boring was concerned, all their enthusiasm, dedication and hard work could not transform the data gathered through testing procedures into a form that would advance psychological science. For him scientific advance was earned via the experimental method.

"The Logic of the Normal Law of error in Mental Measurement," published in 1920, was Boring's most formal criticism of mental testing - some believed it to be his most formidable. His 1919 paper, "Mathematical vs.
Scientific Significance," set out issues that were developed in more detail in the "Logic" paper.\textsuperscript{40} The gist of both papers was that mental measurement merited only a "discriminating encouragement".\textsuperscript{41}

Both of these papers, as well as the ones he wrote later, hinged on the argument that scientific inference can not just be statistical. Ruml would have been sympathetic with such a point of view; his criticisms fit under this broader argument. What was implied was that statistical method can not in and of itself provide a short cut to making scientific generalizations. Statistical method can not replace careful analysis of the problems faced by the researcher. Ruml extended this attitude to research in the applied field; Boring had in mind the problems of mental measurement as they were faced in the laboratory.

Boring was an experimentalist. For him statistics were useful descriptive tools. Statistics did not constitute a full analysis of the meanings inherent in the data. Thus, he cautioned that it was possible for statistical analyses to add meanings to the data. Like all quantitative descriptions, their economy of expression is won through the plausibility of the assumptions that have to be made.

Further to these concerns, Boring thought it was easy to confuse description with prescription, particularly if expediency was important. Thus, although statistics may provide useful mathematical descriptions of the data, they were not in themselves a sufficient basis for making scientific inferences. Boring commented in his 1919 paper that "statistical ability, divorced from a scientific
intimacy with the fundamental observations, leads nowhere."42 He accepted the value of statistical constants as they provided a "conventional method of summarizing frequencies of observed data". He contended however that to "shift the meaning of probability from observed frequency to predicted frequency," was "precarious".43 Boring admitted that science proceeds by attempting such a maneuver, but he urged that "scientific generalization" was "a broader question than mathematical description."44

What made scientific generalization a broader question was simply the conviction that the scientist must engage in an analysis of causes. What caused the frequencies or the variability was the important question for the experimentalist. Statistics could describe frequencies and variability. What disturbed Boring was the manner in which statistics were being used to generate psychological knowledge. For the most part he attempted to demonstrate that any knowledge gained through an uncritical application of statistics was "a-psychological".

Boring's 1920 "Logic" paper was a manifesto setting out the limitations of, and confusions resulting from, the use of statistical assumptions in psychological research. He centered his discussion around three basic categories of analysis: philosophical, mathematical, empirical.45

1. "Philosophical Analysis"

Boring began his paper by discussing the nature of probability. He charged that knowledge can not be wrought from ignorance, and the "problem of probability exists only in the face of ignorance."46 He used the term 'ignorance'
for its value to incite the mental testers. He wrote to Troland that it was his "idea of a bit of scientific propaganda" to equate chance with ignorance and "show it up in all its absurdity, and then confound the group of people that were working at this level." 47

His argument rested on the premise that probability, as such, is never observable. What is observable is the frequency with which an event occurs. Probability, he argued, is assumed on the grounds of observed frequency. He cautioned that it is "plain that the probability of the event can not inhere in the event." Thus, he reasoned, the probability of an event lies in the series to which it belongs. He commented:

When we ascribe a probability to a particular event we are simply seeing that event in a series in which the event is repeated as it varies in some particular phase; or, to put it more picturesquely, we see a series of repeated events telescoped within the single instance.48

These comments sit uneasily with other things that Boring states about probability. From the statements quoted it is conceivable to think that Boring was most attracted to Venn's frequency approach to probability. Yet in his "Logic" paper Boring cited favourably the work of Arne Fisher. Fisher was not a frequency theorist but tended to side with Bayesian interpretations of probability.

That Boring quoted from both Venn and Fisher to support his own position seems odd. Venn was an opponent to Bayesian interpretations of probability, Fisher wrote in its
defense. This apparent paradox did not dawn on Boring. Fisher articulated a particular interpretation of Bayesian theory that appealed to him. Venn showed that we assume probability on the grounds of observed frequency. Both ideas appealed to the experimentalist, both were altered by him.

Boring's position fits neither the views of Venn nor Fisher very easily. This awkwardness in Boring's attempt to articulate a theory of probability is understandable in light of the fact that the modern notion of "Bayesian statistics" was not available to him. What Boring opted for was a special form of Bayesian theory. This development is not easy to sort out.

As an experimentalist, Boring regarded it as the responsibility of the scientist to identify causal factors. For him frequency distributions were determined by underlying causal dimensions. It was a researchers responsibility, he thought, to determine empirically the effects these causal factors had on the frequency distribution of some chosen dependent variable - i.e. a mental test score. Thus Boring reckoned that the proper procedure in science was to first isolate and determine the frequencies of these causal factors. The next step was then to empirically determine the frequency distribution of a chosen dependent variable that results as a function of this "a priori" distribution of causal factors.

This a priori distribution then was not assumed to exist on the basis of the frequency distribution of the dependent variable. Rather, the nature of this apriori distribution
was established on a body of evidence which generated expectations concerning the probability of the events to which they were related (i.e., causally related). If determination of these related events was perfect, then there was no need for probability statements. But causal determination was never perfect and Boring saw 'probability' as an acknowledgement of this knowledge gap. Thus probability was not wrought from ignorance, but was the result of incomplete knowledge of the relationship between causally related events.

Boring tended to be Bayesian in his approach to probability. He was not a "subjectivist" but more of a "logical" or "epistemic" Bayesian. The clarity of his position was variable but I think this reflects the lack of development in Bayesian theory at the time. Michael Acree (1978) noted that the distinction between frequency distributions and probability received little systematic attention. Indeed, he suggested that Venn's book was the most significant attempt to work through the conundrum and he did so by conceiving probability as relative frequency.

Of course there were available other writings on probability that clearly impressed Boring. For example, he was favourably disposed to Arne Fisher's (1915) book The Mathematical Theory of Probabilities. Fisher stated that he supported Bayes and it was only because of misinterpretations of Bayes's theorem that this approach to probability was neglected. He wrote that some of his Scandinavian colleagues had clarified Bayes position. He wrote:
"Unfortunately the rule known as Bayes' Rule has been applied very carelessly, and that mostly by some of Bayes' own countrymen; so the whole theory of Bayes has been repudiated by certain modern writers. A recent contribution by the Danish philosophical writer, Dr. Kroman, seems, however, to have cleared up all doubts on the subject, and to have given Bayes his proper credit.\textsuperscript{51}

Boring borrowed from Fisher his discussions of "cogent" and "insufficient" reasons. These terms fit with the type of "Bayesianism" that appealed to Boring. I will return to these terms shortly.

Further proof of Boring's Bayesian inclinations was his acceptance of John Maynard Keynes' interpretations of probability. Keynes' book, \textit{A Treatise on Probability}, was not available until 1921 and so it was not formative on Boring's early perspectives on statistics. When it appeared Boring was clearly impressed with Keynes' argument. In 1923 Boring wrote to Truman Kelley:

I had more or less the impression that I could retire entirely from the field of the logic of mental measurement since the publication of Keynes' \textit{Treatise on Probability}. He says everything that I wanted to say there so much more fully and effectively and clearly, and with so much greater background, that I suppose the issue might also be considered as settled.\textsuperscript{52}

Keynes was a Bayesian in his interpretation of probability. In discussing the work of Keynes, Acree (1979)
stated that he took probability to be the "degree of rational belief" as warranted by "a particular body of evidence." Furthermore, Keynes interpreted probability as a relation; the probability of a proposition existed only in reference to a body of data. As this body of data changed, so did the probability. Thus the context within which the probability of a proposition was contained, reflected the degree of "rational belief" that the event would occur. As Acree noted, Keynes was not interested in dealing with psychology (i.e. subjective belief) but with epistemology.

Boring's attraction for Keynes is understandable. The 'relation' between the "aposteriori" distribution and the causal "apriori" distribution was conditioned by an independent body of data, a context of knowledge. Actually it was not until 1950 that Boring articulated his position on probability, though one can see these Bayesian ideas in his 1920 paper.

In 1950 Boring circulated a memo through the Harvard Psychology Department. He titled the memo: "Memorandum on Wahrscheinlichkeitsrechnung" and addressed himself as an "amateur statistician with sophisticated worries." In this document, which he never published, he set out his position on the calculus of probability.

What stirred him to circulate his thoughts was a recent book by C.I. Lewis (1946), An Analysis of Knowledge and Valuation. He regarded Lewis as his "Daniel" in his age old battles over the use of statistics. Boring wrote that Lewis defended apriori theory by asserting that "probabilities are not observed frequencies but are inferences from premises."
Lewis reasoned that to make an induction was to "shift from empirical theory to apriori theory." That is, since generalization proceeds beyond the available data, it requires a "credible" data base. Probability, according to Lewis, was a "logical relation" between its "determination and the grounds on which it is judged." On apriori theory, Lewis suggested, "probabilities are determined as conclusions based on evidence". This evidence, the "credible" data base, was akin to Keynes "rational belief". Thus probabilities hold only for particular data.

Boring was attracted to these "logical", "epistemic" interpretations of probability as they required for their meaning a data base that explicitly examined the causal networks among variables. The role of the researcher as he envisioned it - that of an arbitrator of the 'facts' - was maintained on such an interpretation of probability. If the data change, the probabilities change because they are dependent for their meaning on a "body of evidence." Probability statements, and therefore statistical statements, were not a substitute for the judgement of a skilled researcher.

In light of these discussions on Boring's perspective on probability, I will now turn to examine Boring's distinctions between "cogent reason" and "insufficient reason."

Boring was not clear just how he used these terms, as well as others, in his "Logic" paper. He attempted to clarify his point in a letter to F.M. Urban. When he talked about drawing conclusions from 'insufficient' reason he was
referring to the inability to distinguish which event would be most probable and then assuming from this that events were equally likely. Thus the researcher makes "an assumption in the face of insufficient reason and then proceeds to the logical consequences of this assumption." To conclude from cogent reason was, according to Boring, to "determine the objective probabilities first and then to work out the consequences (frequency distributions, etc.) which follow from these observed objective probabilities." What Boring meant by "objective probabilities" was that these events were determined independently through an experimental analysis. In Keynesian terms objective probabilities referred to that body of evidence that provided the grounds of "rational belief".

As Boring stated to Urban, the difference between "insufficient" and "cogent" reason is "whether observation or mathematical explication comes first." In his "Logic" paper he was more polemical than this. He asserted that if one proceeds from insufficient reason, there exists the possibility of fitting the observed data to expected data. Boring believed that this was paramount to acting out of ignorance. Perhaps the observed data would "fit" with the expected, but it might fit other expected models as well. He argued that since observed data would not fit expected frequencies perfectly, and therefore might fit other types of frequency distributions, how would the scientist determine when the data could be considered to "fit" with a priori expectancies. In addition, Boring pointed out, psychologists are always working with samples and it is
possible that the obtained sample just happened to fit with the expected model.

On the other hand, proceeding from cogent reason the psychologist sets out a body of evidence that suggests what the observed frequencies should be and then tests to see if the frequency of an event matches. The obvious problems that crops up of course is how to determine when an observed frequency matches an expected frequency. Boring mentions this problem in his "Logic" paper, but does not offer a solution.58

This set the stage for Boring’s next argument, what he called the fetish for seeing the Gaussian Curve in all of Nature. He wrote:

The Normal law of error has been both an inspiration and a limitation in statistical measurement. . . . The law came to play the part of a first principle of nature, of an ideal, given a priori, to which nature seeks to conform.59

Boring stated that the normal curve was a frequency distribution, one among many. Yet, he pointed out, it had achieved over all other distributional possibilities the acceptance that it was Nature’s preferred frequency form.

Why was the normal curve given such status? Boring attempted to sketch the history that led to this acceptance, citing Quetelet as the one who introduced the curve’s usefulness to the social sciences. In recounting the brief history of the "fetish", he found support for his main point: that scientists have tended to proceed (with respect to the normal curve) from "insufficient reason". Thus
Boring noted that, "Quetelet actually established the law by the data and then corrected the data by the law." For Boring the sanctity did not rest with the normal law but with the data.

The normal curve was assumed to apply to most natural events. Boring could not accept this. He pointed out that although Galton and Pearson originally were caught up in the "fetish" of seeing the normal law in all of nature, they eventually tempered their enthusiasm. So too should psychologists, he thought.

What Boring argued in this section of his paper was that mental testers were working from "insufficient reason" in fitting all their data to the normal curve. They had not established, or even attempted to establish, whether or not the Gaussian curve actually described the frequency distribution of intelligence. Instead of seeking what Boring called cogent reasons for applying the normal curve to mental measurement data, they allowed the normal distribution - like Quetelet had - to shape their data.

Boring's next move was to show that the normal law was not necessarily nature's preferred frequency distribution. Again he appealed to history to make his case. He cited J. Bertrand (1889), W. Lexis (1877) and E. Czuber (1910) all of whom found skewed distributions in social data. He quoted from Raymond Pearl's research on Ceratophyllum which demonstrated that "skewness in variation is a very real biological phenomenon." Thus Boring reasoned in light of such precedents that to assume a priori that the normal law applied to mental data was to proceed from ignorance to
establish scientific truth. He urged that psychologists should not be so interested in the applicability of the normal law as in the facts. "The appeal to facts," he wrote, "is always a protest against theory which is given a priori."61 He noted further that:

When anthropometric measurements were first being made, there stood out the fundamental fact of the massing of cases about the average, the rapid falling off of more extreme frequencies, and the extreme rarity of widely divergent cases. To this extent the normal law was the fact. But as interest centered upon the details, the inadequacies of so simple a generalization became apparent. There was effort enough at reconciliation, but, in general, science kept to the facts, and a more flexible system of representations came into use.62

Boring felt that psychologists, in their enthusiasm to partake of the advantages of a calculus of probability in their research programs, had overlooked a fundamental consideration. He argued that even if it was accepted that nature conformed to the normal law, we would have to determine "nature's own unit of measurement."63 This led to his "mathematical" arguments against the use of the normal curve in mental measurement.

2. "Mathematical Analysis"

Boring noted that Galton in 1879 expressed reservations concerning the application of the normal curve to all data. Galton suggested that if the unit of measurement is altered
so to would be the form of the frequency distribution. Thus, if the psychological dimension was measured in $X$ units and was normally distributed, if one measured the dimension with $X^2$ units, it would no longer be distributed normally. Boring quoted Galton and also remarked that Bertrand drew attention to the same point. He expressed amazement that such a relation that was "so obvious should have received so little attention." 64

I surmise that Boring called this a mathematical analysis because "nature's" deviation from a normal frequency curve was due to a mathematical manipulation. Thus the argument was from a mathematical fact, not from a philosophical point that evaluated the relationship between frequency and probability or from an empirical point that draws on evidence that the normal curve is not always found in nature.

What Boring suggested was that if nature preferred to express its variability via the normal distribution, the scientist would have to discover the measurement units that nature used. If the scientist chose just any unit, and if this unit's measurement function was not linear to nature's, then it would be an error to assume that the normal curve applied to such data. Thus he asked mental testers to query as to whether or not their chosen unit of measurement was linear to nature's unit. He commented:

The case is very different from that of the coin or of the incommensurable number. We do not know anything about the units with which we are working [in mental measurement] except that they are the
units with which we are working. He knew that the mental testers had not explored the measurement functions of their units.

Boring argued that before psychologists proceed with further applications using mental tests, laboratory research should be carried out to determine the relationship between test data and "mental distributions as found in nature." In a letter to Troland he wrote that "to assume that Nature is striving for the normal curve and introduces compensatory variation whenever the correlated factors tend to destroy the normal curve," was simply "not scientific." 66

3. "Empirical Analyses"

In the same letter to Troland, Boring suggested that America is probably the only country where "the Gaussian fetish survives." He suggested that there was a tendency in American psychology to work from a priori assumptions rather than to find out what actually works. In his empirical analyses he drew attention to the importance of not working from such a perspective, what he called proceeding from insufficient reason, and to carry out studies that would empirically evaluate the measurement functions of the units utilized in mental measurement research.

In this vein he found F.M. Urban's work on the distribution of judgement errors in psychophysical experiments to be of value. Urban had demonstrated that they followed the normal frequency curve. Other studies were not so successful in finding the normal curve in their data. Boring cited a study by H.D. Williams (1918) which endeavoured to obtain a psychometric function for memory.
He examined frequency of recall as a function of repetition. Boring stated that although Williams wanted a "truly mnemonic unit", and since he could not find "cogent reasons" for applying the normal curve to his data, he had to "stick to the observed facts, leaving the frequencies to stand as a function of the number of repetitions." He also noted that M.R. Trabue's attempt to establish an equal interval language scale was thwarted because he did not have a real quantitative measurement. Boring suggested that in light of the difficulties faced when equal increments were not empirically evident, Trabue resorted to assume that the normal law held and used the probable error as his unit of measurement. Trabue justified his action saying, "If this assumption is made the results that follow are in . . . close accord with known facts." Boring criticized him saying that he neither provided these "known facts" nor did he suggest the degree of fit.

Boring drew attention to the Karl Pearson and G.A. Jaederholm (1914) studies which applied Jaederholm's version of the Binet scale to mental defectives to establish its graduated character. They did not get a good fit, according to Boring, but concluded that the non-Gaussian character of the distribution of intelligence was owing to the methods of application as well as problems in the test form. They did not conclude that the distribution of intelligence might have deviated from the normal curve. Boring wrote that this conclusion was "as near a confession of what the Biometric School is trying to do as we are likely to get. It is frankly seeking always to see Gauss in
Boring concluded that empirical studies have not always found Guassian curves in nature. Given this circumstance, as well as the fact that an arbitrarily chosen measurement unit may not follow a normal distribution, there was little grounds for mental testers to assume a priori that the normal law should hold for mental measurement. All that there was, Boring suggested, was the "will to believe" in the normal law.

4. Boring’s Conclusions

Boring was secure in the arguments which he put forward in his "Logic" paper. Later when he was asked why he never formally wrote a rejoinder to Kelley's criticisms of this "Logic" paper he said that he had "given up replying". The reason he offered was:

... when I undertook it [the reply] I came to the conclusion that the reply was stated in the article quite as clearly as I was able to state it.

Boring recommended that, "if you read me and then Kelley, I suggest that you then read me again."72

What Boring concluded was that there were not sufficient grounds for psychologists to assume that the normal curve applied to mental data. It was not scientific to proceed in this manner. The units of measurement chosen in mental testing were done so arbitrarily. Although he felt this was a necessary consequence of the research area, he could not condone them being treated as if they were measurements, "in the sense of being the sum of equal units."73 Nor could he accept that it was justified that an arbitrarily chosen unit
would be distributed normally. In light of this consideration, Boring wrote that if it was inappropriate to determine the unit from an assumed distribution, then "we may, nevertheless, determine the distribution from the unit. This, in fact, is the necessary scientific order."\textsuperscript{74}

The matter of the unit of measurement cut deeper. Boring noted that mental measurement did not constitute "a sum of equal units". Yet, he argued, it was clear that the statistics applied to the mental test data assumed that measurement was of this constitution. Mental test data was not considered by him to be really quantitative. He wrote that, "the application of a unit that is not psychological to a quantity that is psychological does not yield a measure of a quantity." What such 'measurement' accomplishes, he suggested, was the "rank-orders of a number of quantities." To use statistics on what is essentially rank ordered data could lead to results that were more "exacting" than was evident in the data itself. Boring recommended that the type of statistics used on mental test data had to reflect the crudeness of the quantitative assessment. "What we must remember," he wrote,

is that we are dealing with the statistics of medians, quartiles, contingencies, and correlation ratios; not with the statistics of averages, standard deviations, coefficients of correlation, and linear regressions. All those statistical constants, that imply a scale of equivalent units, violate in use the conditions of the case and lead to a precision of result that is an artifact\textsuperscript{75}
S.S. Stevens during the 1940s would make this same argument with respect to measurement and statistics and ignite again the debate.  

Boring recognized that there was an attempt to treat mental measurement as "quantity" because physical units—such as time (mental age measured in years), number of items completed, etc.—were used to assess performance in a mental test. But he argued that these physical units may not be psychologically meaningful. That is, to identify physical units with psychological units was a logical error; it was to commit what Titchener called the "stimulus error." According to Boring this was the mistake of behaviorism; it was also the mistake of mental testing.  

For Boring then, psychology must attempt to deal in units which are psychologically meaningful. Although arriving at such units is a painstaking process, once obtained, the psychologist can plot the frequency distributions and proceed to set mental measurement on firm scientific ground. All other ways of securing mental measurement, as demonstrated in behaviorism and the testing movement, were attempts to "a-psychologize psychology." This trend was for Boring a fundamental threat to the establishment of a scientific psychology. I will return to this discussion of the unit of measurement in Chapter Five. It was (and remains) an issue that psychologists differ on in fundamental respects. Thus, the differences between Boring and Kelley over the issue of measurement provides a point of departure for methodological debates that were taken up in the 1930s through the 1950s.
Conclusion: Ruml, Boring and Expert Knowledge

Beardsley Ruml and Edwin Boring agreed that mental testing was proceeding in an unscientific manner. At first this agreement appears to be strange. Ruml was an applied psychologist and was involved in the testing movement. Boring, on the other hand, was an experimentalist and was not favorably disposed to mental testing. Yet their criticisms addressed similar issues. Of course Boring’s articles, being more elaborate, covered more ground. Ruml’s points could be subsumed under Boring’s.

The awkwardness of their agreement rests with the fact that Boring was a pure psychologist and Ruml an applied psychologist. It is commonly reported that there existed tension between applied and pure psychologists. Here we have an example of agreement on a fundamental question of research procedure. What this points to is that the tension between applied and pure research is more complex. Indeed, the tension rests between the acceptability of different knowledge generating practices and not with the means of knowledge application.

Both Boring and Ruml proposed that mental testers had a tendency to overlook the implications of applying statistical methods to mental measurement data. Both critics pointed out the unlikelihood of statistical assumptions holding for such data. They proposed, in different ways, that mental testers should examine the foundations of their practices. They agreed that methods must generate trustworthy knowledge - be it scientific
(pure) or practical knowledge.

Ruml was concerned that applied psychology was being taken for a psychology that would settle for approximation. Boring was concerned that mental testing, in conjunction with behaviorism, were clear attempts to obtain objective measurement scales by "a-psychologizing" psychology. In other words, both mental testing and behaviorism were inclined to assign to physical measurements psychological properties, and ignore any attempt to establish physical measurements that were psychologically meaningful.

Boring's criticisms, as well as Ruml's, suggested that any criterion for the acceptability of a method that eschewed accurate description as its major initiative was totally unacceptable in a science or applied science. They agreed that methods must be explored to ensure that the knowledge generated through them provides an accurate description of the underlying causal variables. Both Boring and Ruml recommended that the place to do this was in the laboratory or a controlled environment.

I would be inclined to say that Boring and Ruml adopted the "objective arbitrator" model of expert. This would not, however, be entirely accurate. Just the same, they both rejected the "managers of methods" expert. Ruml, I think, was actually pulled in both directions. He was an applied psychologist, but he believed in the ideals of laboratory procedure. Basically, he took these ideals and applied them in the work place. In other words, he brought the laboratory into the applied setting. This was the "old" way of doing applied psychology. Mental testing, applied to
masses of people, was the new. The applied psychology that stuck to the ideals of laboratory research, emphasizing the importance of the accurate isolation of causal variables in the work place, was opposed to the applied psychology that compiled mental test data, drew up charts and correlation tables, and proceeded from here to make policy recommendations.

Boring also rejected the notion that statistical methodology could perform the task of isolating causal factors. He reasoned that the adoption of statistical methodology to perform such a task, a task that was generally considered to be the domain of the researcher, was doomed to fail both as a way of doing science and as applied science. The only way such a move could be successful was if the assumptions implied by the use of statistical methods held for the data. Boring believed that this was an empirical question and not something that should be assumed a priori. Laboratory psychology, he thought, was the first step in setting the course of mental measurement. Inference to future conditions could not rest with statistical analysis alone.

Both Boring and Ruml were met with rejoinders from the mental testers. What was at issue was not merely an answer to specific criticisms, but the right of mental testers to generate psychological knowledge through methods that they [the mental testers] believed "spoke for themselves." They argued that their methods were indeed scientific even though they were methods which were independent of the laboratory.
NOTES


9. Thorndike wrote to Kelley saying: "I have an opportunity for you to do a service to science, get recognition for it, and be moderately paid therefore. I judge it will be agreeable to you. It is to make correlation tables of the
distribution of 10,000 pairs when $r = .9$ and when $r = .8$ and correlation is to be of normal rectilinear type." Thorndike suggested that he could publish them separately and simply credit "Carnegie Foundation for aid given". See Thorndike to Kelley, undated, but in the file marked 1915, Harvard University Archives, Kelley Papers. Later Thorndike wrote to Kelley (Feb. 7, 1916) saying that he could not interest the Carnegie Foundation in the publication of the tables, but said that he would "assume the financial risks, leaving the profits (if there should be any!) to revert to you." Kelley preferred to back his own tables and Thorndike pushed their publication through Teachers College Publishers; see Thorndike to Kelley, April 1, 1916, Kelley Papers, Harvard University Archives.


11. H.E. Burtt, "Partial Correlations on a Slide Ruler," Psychological Bulletin (1919) 16 240-242; H.A. Toops, ibid., p. 445-447. Toops wrote: "In efficiency, the combination of
printed sheet and plotting machine has been found through extensive try-out to yield Pearson correlation coefficients in from one-third to one-fourth of the time taken by other methods." (p. 445)

12. Kelley's exchange of letters with the Keuffel-Esser Company are interesting in that it is apparent that Kelley desires to have on the market a piece of apparatus to efficiently calculate correlations. Kelley said that the ruler would replace his book *Tables to facilitate the Calculation of Partial Coefficients of Correlation and Regression Equations*. He argued that since this book was out of print (the edition was 300 copies) and demands to reprint were coming in, he thought the slide ruler would have a market. Keuffel was not convinced. See Kelley to Keuffel-Esser, March 29, 1920; Keuffel to Kelley, April 29, 1920; Kelley to Keuffel, May 12, 1920; Keuffel to Kelley May 22, 1920; Kelley to Keuffel, May 22; and Keuffel-Esser Company to Kelley, July 6, 1920. Kelley did request that the slide ruler bear some connection to his name and that he receive royalties from its sale, (see the letter of May 12).


18. The assumptions Ruml was aluding to were assumptions such as the normal distributions, equal interval scales, and that linear relationships - rather than curvilinear ones-held between dependent and independent variables.

19. Ibid., p. 59.

20. Ruml to T.L. Kelley, January 5, 1920. At time of writing this letter Ruml was employed with the Scott Company (as stated on the letter head, "Consultants and Engineers in Industrial Personnel"). Harvard University Archives, T.L. Kelley Papers.

21. Ibid., p. 60.


24. Ibid., p. 466.


26. Ibid., p. 184.

27. Ibid., p. 184.

28. Walter V. Bingham quoted several complete passages from this paper in "On the Possibility of an Applied Psychology," Psychological Review (1923) 30 289-305, see pp. 301-302. I have not been able to gain access to the original document. It should be noted that Ruml was Bingham's assistant just prior to the War.

29. Ibid., p. 301.

30. Ruml, "The Need for an Examination; etc.", op. cit., p. 58.

31. This passage was quoted from Martin Bulmer and Joan Bulmer, "Philanthropy and Social Science in the 1920s," op. cit., p. 366. It is cited by them from Ruml's "General Memorandum", pp. 21-22.

32. E.G. Boring to Raymond Pearl, August 30, 1920. Harvard University Archives, E.G. Boring Correspondence.


35. Ibid., p. 28-29.


37. Boring, Psychologist At Large, op. cit., p. 31.

38. Ibid., p. 31.

39. I mention this only because I have come across this quote from Boring in other research on the history of mental testing and it has been used to indicate that he supported the research of the mental testers. Most recently, see Russell Marks, The Idea of IQ, Washington DC: University Press of America, 1981, p. 48. I should add that Marks' work, in other respects, is excellent. I have found it to be a useful analysis of the testing movement, though I disagree with him on this point. Perhaps in later years Boring was not so harsh in his criticisms of mental testing as he was equally concerned with the moves in experimental psychology to adopt statistical methods and use them inappropriately.


41. "The Logic of the Normal Law of Error in Mental Measurement", American Journal of Psychology (1920) 31 1-33, p. 1. Here after I will refer to this paper as "Logic".


45. Boring did not set out these categories in his 1920 "Logic" paper but he did in a letter he wrote to Troland in the hopes of clarifying the issues. I have adapted them from this letter as I think they help to organize Boring's detailed arguments. See E.G. Boring to Troland, Dec. 6, 1920. Boring Correspondence, Harvard University Archives.

46. "Logic", p. 4.

47. Boring to Troland, December 6, 1920. Boring Correspondence, Harvard University Archives.

48. Ibid., pp 6, 7.

49. I have drawn this distinction from Michael Acree's excellent Ph.D. thesis (Clark University, 1978) titled, *Theories of Statistical Inference in Psychological Research: A Historico-Critical Study*. Acree defined "logical" or "epistemic" Bayesian interpretations of probability as those that regard probability as a "degree of rational belief". He wrote that it referred to the "degree of belief in a proposition objectively warranted by the data." Acree suggested that the "logical" theorists "try to make probability relative to the whole body of available
evidence." (p. 189) He defined "subjective" Bayesian probability as "personalist", as interpreting probability as "'your' degree of belief in a proposition." Acree commented that "subjective theorists impose sufficient constraints to make everybody's probability appraisals coincide." (p. 189) In this way he sees a merging between the logical and subjective Bayesians.

50. Ibid., p. 190.


52. Boring to Kelley, February 27, 1923. Boring Correspondence, Harvard University Archives. Kelley replied saying, "I cannot agree with you that Keynes has settled this question." (March 13, 1923)

53. Michael Acree, Theories of Statistical Inference in Psychological Research: A Historico-Critical Study, op. cit., p. 192. Acree's excellent thesis provides a clear and readable discussion of Bayesian theory and offers a good discussion of Keynes ideas. The following discussion relies to a large degree on Acree's interpretation of Keynes's theory.

54. Ibid., p. 194.

56. C.I. Lewis set up the situation as follows: "That ω, having a property \( \psi \), will also have a property \( \phi \), is credible on data \( D \), with the expectation \( a/b \) and the reliability \( R \)." See Lewis, *An Analysis of Knowledge and Valuation*, p. 305, quoted by Boring in his "Memorandum". Without going into detail on the "reliability" of a probability judgment, Boring suggested that it was dependent on three basic points: "(1) on the adequacy of the data, (2) on the proximateness of the data (how closely do the instances fit the reference class?) (3) on the uniformity of the data (how much alike are observed frequencies for different samples?)." He drew these basic points from Lewis, but they fit with his orientation to interpreting probability statements. They all involved the judgment of the researcher and there was no room for scientific inferences to be determined by statistics alone.


58. See "Logic", p. 16, footnote # 46.


60. Ibid., p. 14.

61. Ibid., p. 17.

62. Ibid., p. 17.

63. Ibid., p. 15.

64. Ibid., p. 18.

65. Ibid., p. 23.
66. Ibid., p. 21, footnote #55.

67. Boring to Troland, Dec. 6, 1920. Harvard University Archives, Boring Correspondence.


70. Pearson and Jaederholm, "Mendelism and the Problem of Mental Defect, II: On the Continuity of Mental Defect."; "Mendelism and the Problem of the Mental Defect, III: On the Graduated Character of Mental Defect." Boring cites these papers, ibid., p. 24, note #67. The are taken from Pearson’s *Questions of the Day and Fray*, essays 8 and 9.

71. "Logic", p. 25.

72. Boring to O.W. Richards (a lecturer in Zoology at the University of Oregon), March 3, 1924. Harvard University Archives, Boring Correspondence.

73. Ibid., p. 28.

74. Ibid., p. 30.

75. Ibid., p. 33.

Boring's criticisms of mental measurement had some impact. Although his views were not adopted directly in the published literature, he received commendations from several important psychologists and statisticians. In 1920 Raymond Pearl, the noted biometrician at Johns Hopkins University, wrote that he was "delighted to read" Boring's short note in *Science* on Ellis Michael's (a biologist) misuse of statistics. Pearl went on to express that this "flaying" was "much needed" and thought it was "splendid" that Boring was "going into this field in a thorough going way." He requested a reprint of Boring's "Logic" article and they began a correspondence.

This support by someone who was considered to be a superior statistician was taken as a high compliment by Boring. Other respected psychologists, such as Godfrey Thomson and F.M. Urban, agreed with much of what he set out in his articles. After reading the "Logic" article, Thomson wrote saying that: "as far as a first reading carries me I can agree to all that you say in the concluding paragraph." Boring replied that it was "extremely encouraging" to receive Thomson's support and that he was pleased that they were not "working at cross purposes."

F.M. Urban, a psychologist who had written some important papers on psychophysics, wrote saying that he spent much of his spare time studying Boring's "Logic" paper. He claimed that he did not want to write to Boring until he had "mastered" all of the points. Urban commented:
It seems to me that you must have been troubled by just about the same problems I was, for otherwise our ideas hardly could coincide to such a remarkable degree. Some of your passages are such that I am sorry they were not written by me.\textsuperscript{5}

Urban’s letter, which was 15 pages long, was not entirely supportive; it led to an exchange of views that highlighted their differences. Nevertheless, on the main points and certainly in spirit, he agreed with Boring. They used the correspondence to achieve convergence on their ideas concerning the application of probability & statistics to psychological data.

Boring appeared confident that he had issued a blow to the uncritical use of statistical methods in psychology. No psychologists came forward immediately to criticize his articles. Ellis Michael did, but Boring considered this rejoinder to be of little consequence.\textsuperscript{6} He related the story to Raymond Pearl that Michael had taken many of the ideas for his reply from their recent correspondence. "He rather played the role of pupil," Boring wrote, "and he is a good pupil for he is now showing what he has learned from our correspondence."\textsuperscript{7}

Three years later, in February of 1923, a psychologist responded to Boring. The reply was somewhat unexpected. The psychologist was Truman Lee Kelley and his criticisms of Boring were substantial. Before turning to these criticisms, it is important to gain some sense of Kelley’s background and ambitions.
Truman Lee Kelley (1884-1961)

Truman Kelley was one of psychology’s most eminent statisticians during the 1920s. He received his Ph.D. in 1914 from Columbia University under the direction of E.L. Thorndike. He taught briefly (1914-1917) at the University of Texas and on April 20, 1917 he accepted an offer from Teachers College to join its faculty. He served in the War effort as a statistician working on the standardization of the trade tests. In 1920, at the request of Lewis Terman, Kelley joined the faculty at Stanford - appointed jointly to education and psychology. He remained there until 1931 when he accepted a position at Harvard’s School of Education.

Kelley served on numerous committees and in 1921 he was invited by the statistician H.L. Rietz (University of Iowa) to join the NRC Committee on the Mathematical Analysis of Statistics. Raymond Pearl was also a member of this committee. In 1926 Kelley was appointed vice-president of the American Statistical Association and in 1938-39 he was president of the Psychometric Society.8

Kelley’s work in statistics was well thought of by some eminent statisticians. Ronald Fisher wrote to him after reading his text Statistical Methods (1923) saying that it was "quite the most useful and comprehensive books of the kind yet written."9 Pearson was less complimentary. He thought that "much of the book [was] good," but since he disagreed with Kelley on certain points he declined reviewing it for Biometrika as he could not give it an entirely positive review.10 The psychologist Cyril Burt wrote to Kelley in 1944 suggesting the importance of
Kelley's text. He wrote: "After twenty years, Statistical Method is still the book on the subject for every psychologist in every teaching department."^11

It is likely that Terman wanted Kelley at Stanford principally because he needed a good statistician. As Terman remarked in his autobiography, he was lacking in statistical training as Clark (where he did his PhD.) did not offer courses in the area.^12 Later he managed to work with Thorndike (1916) but he always remained aware of his lack of statistical understanding. Indeed, when Kelley favourably reviewed his Measurement of Intelligence (1916), Terman wrote to him: "I appreciate it more than I can tell you, although I realize very keenly how much less adequately the data have been treated than would have been done by one of your statistical expertness."^13 On September 3, 1920 Kelley accepted the position at Stanford.

Terman thought highly of Kelley. In June of 1950, Terman gathered together Kelley's former colleagues to honor him. Terman said that Kelley was "among the half dozen psychologists of the world who have done most to place the new science of psychometrics upon a sound basis."^14 Not only did he appreciate Kelley's statistical know-how, they worked well together. Terman wrote to Boring when Kelley accepted Harvard's offer that it was a "great blow" to loose him.^15

Kelley's perspective on statistics reflected the growing trend to find in this methodology something more than concise descriptions of data. Certainly as an undergraduate the works of Pearson and Galton were
presented to him in the context of hereditarian theory. It is clear from his undergraduate notes that he was aware that Galton and Pearson employed statistics to extend their hereditarian notions. Yet, according to Kelley, it was not until he started writing *Statistical Methods* that he began to see that statistics provided a tool from which one could build theory. In a letter (1926) to Helen Walker he commented:

> It was not until I was deep in the process of writing "Statistical Method" (ca. 1919-1920) that statistics became a powerful device for discovering issues. I now feel that the person who merely uses it as a tool in the development of hypotheses otherwise derived has scarcely a glimpse of its possibilities.

He suggested that statistics was as "superior" a method for arriving at hypotheses as "inductive logic [is] to deductive."17

Perhaps it was Kelley's visit to Pearson's Biometric laboratory in 1922 that stamped in the notion that statistics were useful in generating theory. He recalled this visit as a landmark experience. He credited Karl Pearson as being "the greatest single influence" in his own development as a statistician.18 Kelley added that he wished he had visited Pearson before writing *Statistical Method* (which he said occurred between 1919-1922) as "some of the shortcomings of that work" would have been avoided.19

Kelley, like others interested in mental testing, had an
interest in developing socially useful institutions that depended on testing expertise. He, like Terman and Thorndike, had the idea that the qualities of individuals could be discovered through testing procedures and the application of statistics to such data. They all shared the conviction that such research was of the utmost social importance. From early in his career he was involved in promoting the proper use of mental and vocational tests. In January of 1920 he proposed to Dean J.E. Russell of Teachers College, that they develop a personnel clinic, "similar in some respects to that undertaken by the Carnegie Institute of Technology and by the Scott Company." He suggested that it "could be kept upon a strictly professional basis, and would tap almost unlimited wealth, because of its appeal to men in big business." He added that he thought that this type of endeavour was an appropriate "function of Colleges of Education."20

Further demonstration of Kelley's activities on behalf of the mental testing community, as well as his commitment to the social importance of their research, can be found in a letter he wrote to the government Commission for the Feeble Minded, chaired by Dr. Bailey. He urged in this letter that the passage of a Bill, as outlined in Legislative Document 44 - which was setting out guidelines for the diagnosis and care of the feeble minded - would be a "retrogressive step". His complaint was that the bill had not been explicit in "designating the technical qualifications" of those who were to administer such programs, and there was no outline of just how these
administrators were to be chosen. Kelley advised that there was an "urgent need" that only "technically qualified" individuals be allowed to diagnose, classify and educate special groups. He stated further, and this certainly carried the brunt of his point, that as he read document 44, "any physician is assumed to be competent to pass upon all these questions." 21

Kelley thought that psychologists were eminently qualified for such socially important tasks. This did not lead him however, to think that psychologists' methods of individual and group assessment could not be improved. Rather, it was his goal to develope better, more efficient, methods of evaluation. His proposals for research funds reveal the means through which this development was to proceed. These proposals are full of statements about how statistical methods allowed for the development of procedures that would lead to better assessment practices.

For instance in July of 1922 Kelley submitted a proposal to the Commonwealth Fund. He was in England at the time working with Pearson. Since Thorndike and Cubberley were on the Board he sent them letters concerning the proposal. He wrote to Max Farrand, who chaired the Fund, that the general purpose of the study was to "provide a simple and realistic means of determining the special fortes and weaknesses of a child's intellectual equipment, with a view to utilizing this knowledge in the better education of the child and particularly in enabling a more rational counseling in his prevocational school period." The means of determining these special traits had to reflect "reliable
measurements" which would be determined through statistical analysis. Also he proposed to use the study to test, through an examination of the intercorrelations among traits, whether Spearman, Thorndike or possibly Thomson were correct in their theories of mental structures. He added that he had already begun work on the project in "Professor Pearson's laboratory."22

After submitting the proposal Kelley wrote to Thorndike concerning some specific aspects of the study. He suggested three "major problems" that he hoped, "with Pearson's help," to solve. He listed them as:

(a) the bearing of type of distribution upon trend of evolution. (b) the statistical procedure for solving the following problem: Given measures of n traits of a number of individuals. Required, the determination of such sub groupings of traits as are most independent of each other and most descriptive of the individuals. . . (c) the determination of the equations of growth curves of different mental functions, yielding evidence especially as to saltatory periods.23

The interesting point in this quote is that Kelley saw statistics as the primary tool for answering these questions. Also, there is a sense here that he hoped to reject Spearman's structure of the intellect. These statements show Kelley's acceptance of a more "Thorndikian" model and his hope to develop equations that would demonstrate this model to be correct.24 His full working out of this problem, and his attack on Spearman's model, was
published in *Crossroads in the Mind of Man* (1928). This book was influential in American psychology and set off his controversy with Spearman and Spearman's students.25

Thorndike replied that the Commonwealth fund might find the study to be too "psychological rather than educational" and refer it to the "Carnegie Institution." He suggested that Ruml had recently taken over control of the Laura Spellman Memorial and he may have more success there with such a project.26 Kelley stayed with the Commonwealth Fund.

His study was initially rejected. Kelley wrote to Farrand highlighting the importance of such a study for practical considerations in education and vocational guidance. Farrand replied that the "restatement of the problem . . . put it in a different light" and being that the project was of interest, they would consider funding it at their next meeting. This meeting was held in the fall of 1923. Cubberley sent Kelley a telegram: "At Commonwealth Fund Meeting just closed finally succeeded in getting you grant $7500 for your study." The telegram also noted that Thorndike was able to "pry up" the award.27

So far I have attempted to demonstrate that Kelley was a significant member of the group of mental testers. They shared research goals and were supportive of each other. In the proposal related above, Thorndike and Cubberley undoubtedly used their influence to gain the award for Kelley. They shared a conviction as well that statistical methodology not only provided a means of working out the best theories for mental structures, but led to better applied practices.
Kelley's ability to use and develop statistical techniques made him a respected member of this group of mental testers. This can be seen in Thorndike's, Cubberley's, and Terman's correspondence with him. For instance, when Thorndike petitioned the Carnegie Corporation to establish what he called a Unit Traits Committee (UTC) which would plan investigations into "the unitary differential traits of human nature" he sent a draft of the proposal to Kelley. He added a confidential note:

"You must do this, being absolutely indispensible. If we do a good job with the $5000, we will get $5000 more to perfect the plan, and then may get $250,000 to operate it."28

In keeping with this, Gail Hornstein (1984) noted that the members of the board for the development of the School Intelligence Scale--meeting in March, 1919--agreed to hire Truman Kelley as their statistician. She stated that they set him to work on calculating correlations on the various tests and subscales so that they could decide on which ones to ignore in the development of their new scale. Hornstein commented that: "Thorndike urged that the results of Kelley’s analysis should serve as the sole criterion of a test's inclusion." Further to this point, Hornstein noted that the board generally felt that they could vote on these matters through the mail. I agree with Hornstein that these events indicate that the board "perceived no need for extensive discussion of Kelley’s recommendations."29 They trusted both Kelley and the ways he employed statistical methods.
Defending Mental Measurement Methodology

As the use of statistics became more popular, debates concerning their use joined psychology’s periodical literature. As statistics became acceptable to more psychologists, the level of debate over their application grew more sophisticated. This led to a situation where fewer psychologists could participate in the discourse of controversy even though more researchers had more resting on the outcome of such debates.

Kelley defended the use of statistics. In the early 1920s the criticisms of statistics drew upon arguments as to the limited applicability of such methods to psychological data. Kelley addressed these criticisms by showing their broad applicability and their usefulness as an inductive technology. By the end of the decade, the ‘outside’ critics gave way to critics from within the ranks of mental testers. Kelley became embroiled in discussions that involved differences over technical statistical matters, such as the calculation of probable errors. In the sections which follow I focus on Kelley’s response to E.G. Boring and Beadsley Ruml, both of whom belong to the former group of critics.

Addressing Ruml

Kelley’s first defense of the use statistics was in response to Beardsley Ruml’s (1920) criticisms. As I outlined in the last chapter, Ruml inveighed against the indiscriminate use of statistics in mental testing research.
Kelley co-authored the rejoinder with Lewis Terman. The strategy of their argument can be stated simply: the use of statistical methodology is in practice a procedure that is as 'scientific' as traditional research methodologies. In keeping with this, they concluded their paper saying that the "mental test method" had "become the most important method of experimental psychology." This was a theme Terman championed later in his Presidential address to the APA in 1923.

Their rejoinder to Ruml was broken down into two distinct parts. The first part, written principally by Kelley, dealt with Ruml's statistical criticisms. The second part spoke in general about the history of science and in particular about how mental testing fits in with the general progress a science makes when new methods are introduced.

Recalling from Chapter Three, Ruml perceived mental testers to be "dominated by the desire" to use correlational and other statistical techniques. This led, he thought, to a lack of concern on their part to violations of linearity. His objection followed two lines of criticism. The first was that there was no reason to assume that intelligence was a linear concept. The implication of this was that there was no reason to add up scores across subscales and use this sum as an indication of general intelligence. The second objection to the linear assumption was that "true" intelligence was not likely to be regressed linearly on test scores.

Ruml used a metaphor to set out the terms of his first
objection. He argued that summing across subscales on an intelligence test was like taking the height and weight of individuals and adding them together to indicate a person's size. Thus he suggested that a tall, slim man may have the same summary score as a short, fat man. Likewise, Ruml suggested that intelligence was made up of a variety of qualities and that summing across them led to a quantitative evaluation that was misleading.

Kelley & Terman's response to Ruml's first criticism can be summarized by one of their statements: "Concepts of aggregates do not preclude recognition of detail." Their answer included two basic points. The first was that they agreed that intelligence was not a unitary trait but rejected Ruml's example of height and weight as an illustration of the point. Instead Kelley and Terman suggested that the relation between the subtests of an intelligence inventory was more a kin to "hip height" and total height.

What was implied - at least it is possible to read their response in such a light - was that an underlying factor for "height" could be found in "hip height". That there would be a positive correlation between hip height and overall height was likely. They sought to strengthen their contention that hip-height and total height was a more pertinent example than Ruml's when they stated that: "Spearman could not have made out as good a case as he has for a 'single mental function' if the usual correlation between mental traits were not high." So what Kelley & Terman objected to in Ruml's example was his definition of
qualitative differences; his metaphor lessened, or obscured, the meaning of the correlation between subscales.

Kelley and Terman therefore did not object to intelligence being regarded as multidimensional. They required, however, that the different dimensions be seen as related. Ruml's notion of the "size of a person" as a metaphor was too gross a concept. Kelley & Terman put forward 'height' as both a more manageable and meaningful metaphor. Height could be broken down into subcategories of leg length, hip length, back length, neck length, etc. All of these shared a common property of length although they were qualitatively different parts of the body. Of course each quality would have a different correlation with overall height of a person, some being more predictable of overall height than others. They also realized that a summary of each dimension to form a composite score functioned merely to rank order individuals.

Kelley and Terman suggested an additional metaphor to bring home their point. They wrote that if summing across subscales resulted in measures of intelligence that had "richer interpretive value" than the scores taken separately, then such a procedure was justified. They argued that just as a mining engineer secures samples that speak to the value of a mine - this value being expressed in "linear dollars" - so too does the tester express intelligence in linear terms. They thought that if the purpose of the mining evaluation was different, i.e., buying mining equipment, then a detailed analysis of the nature of the core deposits would be required.
It is clear that Kelley and Terman conceived of the use of linear scales not so much as an attempt to "measure" intelligence as to "evaluate" the level of an individual's intelligence with respect to a group. The assumptions of linearity allowed for a serviceable rank ordering rather than a quantitative measure. This was a substantial move in the direction of an operational psychology. This shift is important and I will provide a fuller discussion of it in the next chapter.

Ruml's second criticism, that intelligence was not a linear function of test performance, was handled by Kelley. Ruml wrote to Kelley that the prevalence of non-linearity in trade test question relationships rendered correlational methods inappropriate for data analysis. He also noted that he knew of no methods which would "produce proper rectification." "[T]he adjustment of scores to make correlation technique available for use", Ruml added, testified to the point that mental testers wanted to use correlation methods and would create "derivative facts" to do so.37 Ruml continued his letter:

Our thinking has been so dominated by the desire to use correlation and the other statistical techniques that we have been willing to make assumptions uncritically and to distort our data unjustifiably. Until we are able, from a statistical point of view, to approach our raw data with sophisticated naivete, committed to no particular technique and ready to admit that in many cases no technique is available, the interpretive contribution of mental test work
will be about what it has been in the past.38

Kelley accepted the evidence that in trade tests there was a frequent departure from linearity. He denied Ruml's allegation that mental testers had a disregard for such evidence because they were committed to using correlational techniques. "I do not care whether we use the partial correlation coefficient or not", Kelley wrote, "but I would say that it is necessary for the most reliable results to approach raw data committed to a particular principle." He stated this principle as: "elements must be evaluated in terms of their contribution, independent of the contributions of other elements."39

Kelley argued in his letter that this proposition was "defensible absolutely independent of rectilinearity or normality." He reasoned that science must isolate the effects variables have on each other. He realized that in the laboratory the causal relationships between variables were established through controlling conditions. In the applied setting, he thought that the only way to isolate the unique effects variables had on one another was through calculating partial correlation coefficients. In the concluding paragraph of this letter, he wrote:

Let me say that I sincerely hope you may be able to work out a technique that will evaluate the significance of dependence in non-rectilinear data, enabling the handling of the problem with neat facility that partial correlation does in the case of rectilinear relationships. My own devices are, at present, to build up rectilinear functions of the
raw non-rectilinear data and then proceed by partial correlation.

Kelley was not as candid in his published response to Ruml. He suggested that it was "extraordinarily rare" to find anything other than linear relationships in mental tests. On those "rare" occasions when non-linearity was encountered, Kelley suggested that some researchers were on the way to developing techniques to handle the problem. He wrote: "The building up of a technique for handling problems of this sort merely awaits the need for such." The implication was that the need was not pressing at this time.

The remainder of the paper by Kelley and Terman addressed some of the more philosophical aspects of Ruml's attack. They rejected his approach which they portrayed as placing the definition of terms as the first order of business. Ruml had objected to attempting to measure intelligence when there was no agreement concerning the concept. One year later, in a statement written for a symposium on intelligence and its measurement (1921), Ruml commented that the nature of intelligence could not be debated as there was a "lack of precision in the terms and concepts that must form the basis of such a discussion." He also noted that there was an "absence of factual material" which could speak to some "essential points."

Kelley and Terman advocated that definition was not the "essential genius which leads to discovery". Rather it was through methods that truth about individuals was revealed. This led to discovery. They wrote that if something exists, "as a phenomenon of human nature and can be measured, then
it is entitled to a name and a definition in terms of its experimental setting no matter if it cuts athwart long established concepts." Ruml's procedure to first define the construct and then choose methods suitable to the task was seen as doing things backward.

The operationalism of Kelley and Terman reflected a new turn in psychology's methodological development. It was a reorientation that gave priority to the methods that made psychological phenomena visible. These phenomena were seen as "facts" and they believed that it was these "facts" that had to be contended with, not their definitions.

Ruml's objections of course were not merely quibbling over proper definitions. He objected to mental testing "facts" as meaningful aggregates. Basically, Ruml required that before data can be combined there has to exist a theoretical rationale that draws together the different components that make up the aggregate. Kelley and Terman, on the other hand, saw in statistical methods a meaningful way to create aggregates of psychological phenomena.

Kelley's arguments were further refined in his response to Boring's papers. Though the discussion between Boring and Kelley became more involved, many of the points of dispute were similar to those that were addressed in the debate with Ruml.

**Addressing Boring's Criticisms**

Kelley was the first mental tester to publically respond to Boring's 1919 and 1920 criticisms of the use of statistics in psychological research. This response came
some three years after the publication of Boring's "Logic" paper. In February of 1923 Kelley sent a pre-publication draft outlining what he thought to be the short-comings of Boring's strictures against the use of statistics.

Kelley's delay in responding to Boring's critique presents a problem. Kelley was writing Statistical Methods between 1919 & 1922 which, no doubt, demanded much of his time. Still, he managed to respond to Ruml. Perhaps Kelley wanted to reply earlier - and there are some indications that he told Boring he was planning a response - but he obviously experienced set backs in getting around to addressing Boring's arguments.

Kelley likely wrote his response while he was visiting Britain or shortly thereafter. Perhaps his entrenchment in Pearson's biometrics laboratory presented him with a working example of the possibilities statistical analysis held for the social sciences. He had claimed that it was during the writing of Statistical Method that he began to see statistics as a "powerful device for discovering issues". Certainly his time with Pearson would have provided further confirmation of his growing conviction.

Also, as MacKenzie (1982) pointed out, the Biometric School was a close knit group and critical of those who did not see the advantages of their methods. Kelley noted in a letter to Cubberley:

Pearson is methodical, keen, and like a razor in the sharpness of the distinctions that he makes and like an Indian in the memory he has for the professional faults of others. The trouble with Pearson is that
he is right in his charges of error on the part of others -true a man may do 99 things correctly and one thing incorrectly but if so the critic is "right" in calling the one thing wrong.

He went on to say, however, that Pearson "is attacked in the same way that he attacks."46

It is likely that Kelley discussed Boring’s criticisms of statistics with others in the London laboratory. Even though it is difficult to demonstrate whether or not these individuals had a direct effect on what Kelley wrote, it is reasonable to believe that he would have discussed Boring’s arguments. Also, given the importance to the Biometric School of redressing critics, Kelley may have gained some inspiration from his British colleagues.

Whatever the case, Kelley did not reply to Boring until February of 1923. The reply was sharp. "Boring’s conclusions are generally destructive " he wrote, "and tend to leave one with the feeling that there is no sound statistical basis for mental measurement, and little for other psychological measurement."47 He suggested that Boring did not possess a good understanding of the statistical practices he criticized. Further, Kelley stated that statistics were much broader than Boring pictured.

Kelley divided the article into two parts; the first being a critique of Boring, the second being a discussion of mental measurement as it pertained to units of measurement. The second part of the paper was also a criticism of Boring, though more indirect. Kelley objected to the way in which Boring discussed the ‘unit of measurement’ and thought that
he made a "material mathematical error" by relating the psychometric function to a function of the unit of measurement.\textsuperscript{48} I will return to these discussions of the unit in next chapter. For now I want to consider Kelley's direct criticisms of Boring's papers.

Kelley began by setting out his criticisms of Boring's (1919) paper, "Mathematical vs. Scientific Significance". He described, through an example, the logic of the Chi-square statistic. The point of the exercise was to present which assumptions were involved in the "derivation of the measure P". Kelley wanted to bring these assumptions to the surface to show that they were the very ones that Boring criticized in his articles.

Kelley wanted to make the point that insomuch as Boring used "P" to describe his data, he was implicitly subscribing to the same assumptions that supposedly impeded the statistician.\textsuperscript{49} In this way he was attempting to disclose the inconsistencies in Boring's position. Furthermore, he wanted to show that good science proceeded by making these types of assumptions whether or not scientists were aware of them.

What was this "measure" P? Boring noted that although the ratio of the difference between two measures to the probable error of the difference was not a "direct measure of significance", these ratio values could be translated into a "scale of 'probability of difference' by the use of a table of the probability integral." This scale ranged from zero (when there is no difference between measures) to "unity, when the difference is infinite with respect to its
P.E." He then said that this scale is sometimes talked about as the "probability that the difference is not due to chance." Boring also noted that it was used (in the case of Chi square) as a "measure of homogeneity and heterogeneity". He also understood Pearson to use 'P' as the probability that the deviation of one curve from another is "random."50

Kelley asked the question: "What assumptions are involved in the calculation of P?" His answer: one. He chose to express his understanding of 'P' through the logic of the Chi-square statistic. He wrote:

... if the obtained series is in reality a random sampling of the series given by the standard, then the random distribution of each cell frequency around the standard cell frequency as mean is given by the normal law of error.

Clearly the important component of determining P rested with the assumption about the normal law of error. "It is assumed" he continued, "that successive random samplings" of the same basic data as represented by the standard would yield sometimes more and sometimes less than the expected value. The expected value would be the mean and the distribution would be normal.51

Kelley's point was that if one used a "P" value, one implicitly assumed the normal distribution. With this point he tried to show that Boring's arguments against the use of statistics were inconsistent. Kelley reasoned that although Boring railed against assuming the normal distribution, at times he recommended using "P".

Of course Boring did not perceive an inconsistency as he...
did not accept "P" as a solitary index for making scientific generalizations. Thus he chose a title for his first article that explicitly set out to contrast scientific and mathematical inference. He wrote:

It appears that the apparent inconsistency between scientific intuition and mathematical result is not due to the unreliability of professional opinion, but to the fact that scientific generalization is a broader question than mathematical description. In scientific work we deal with samples, whereas we are always interested in the larger group of which the samples are intended to be representative. The mathematical formulae do truly measure the difference between the particular samples observed.52

Boring's apprehension with respect to the measure "P" was not with its statistical characteristics but the interpretation of these characteristics given that the numbers that go into the statistical formulae are based on data from a sample.

Boring was willing to admit that if the sample was representative of the population, then the statistical measure "P" could be meaningfully employed. His concern was with the question of how does one assess whether or not a sample is representative of the population when the population parameters are unknown. Under circumstances where one really does not know the representativeness of the sample, Boring argued that the measure "P" was misleading.

Since no statistical methods existed for determining the
representativeness of the sample with respect to the population, Boring argued that one had to go outside of statistics. "[S]ince in the nature of the case," he wrote, "it is impossible [for the scientist] to state in numerical terms the degree of representativeness that his samples possess, conclusions must ultimately be left to the scientific intuition of the experimenter and his public."\(^5\)

Scientific inference, as far as Boring was concerned, involved judgments that included information other than just statistical statements which employed the "measure P". Only when the conditions under which "P" was meaningful were met could such a statistic be used in an inferential capacity.

Kelley rejected Boring's entire argument. He interpreted Boring as suggesting that the scientist must hold to theoretical convictions in the face of contrary data. That is, if one finds in the data a "significant difference" between groups, Kelley argued that this was an observed "fact". To interpret this observed difference as not being a "fact", flew in the face of scientific procedure. Kelley commented:

If we "feel inclined" to draw a conclusion at variance with the data, that settles the matter... If he will not trust such mathematical findings as are contrary to his wishes, that fact in no sense provides scientific warrant for discarding them.\(^5\)

Kelley regarded the scientific procedure recommended by Boring as appearing to "give warrant for keeping and using data if they support an established conviction and otherwise discarding them."\(^5\)
It is clear that Kelley and Boring did not agree on which data should be considered to be theoretically important. "Mathematical findings", though they were considered by Kelley to be theoretically important, were not so for Boring. Formulae could not be a substitute for scientific judgement. Theoretically pertinent "facts" for Boring had to first meet specific analytical criteria that resided outside the realms of descriptive statistical analysis.

This rift was identified by Kelley as reflecting Boring's deductivist interpretation of statistical methods. Kelley, as I have already indicated in previous chapters, saw statistical methods as inductive. Statistics were used as a means to "automatically" generate theory from data, as Muliak (1987) has so aptly described such practices. These methods were used to supplant the analytical processes of the psychologist in discovering the latent components that gave rise to the visible data. In Kelley's program, psychologists had to first be the masters of methods and manage them appropriately so as to generate theory.

Boring, on the other hand, was more continental European in his research perspective. He regarded data as bound by *apriori* assumptions. Science was a process of breaking down these assumptions (analysis) and then proceeding through synthesis to reconstitute the correlative networks that resulted in the visible data. Later in the only work where he was strictly theoretical, Boring (1933) wrote that he was "convinced" that it was wrong to consider that a "strict opposition" existed between "observation and speculation."
He stated that it was a "naive epistemology" that held that "science espouses observation and rejects speculation." He went on to write:

Actually this whole matter can be regarded as a question of the use of hypothesis in science, and it seems to me that there cannot be any longer a doubt that profitable observation must be predetermined, as to the nature of the correlation which it seeks to establish, by hypothesis. The valid dichotomy lies between useful hypothesis and dangerous speculation, and here the line of demarcation is necessarily indeterminate and personal.\(^5^7\)

In Boring's view, scientists had to develop craft knowledge, train their intuitions, serve an apprenticeship under a skilled scientific practitioner. Scientists, he thought, must reason about their data and not have methods - be they statistical or otherwise - supplant the analytic processes that must precede data collection and constrain interpretation of it once it is collected.

It is clear in Kelley's debate with Boring that he regarded a statistically significant result as a "fact" with which a psychologist must theorize. A statistical finding demanded theoretical attention. Thus statistical method provided the data - as did laboratory based methods - upon which psychological theory could be constructed. Statistical significance was scientific significance.

It is interesting that in Kelley's rejoinder, he reversed many of the key phrases that appeared in Boring's articles. For example, Boring wrote in "Mathematical vs.
Scientific Significance" that "scientific generalization is a broader question than mathematical description". Kelley said he would be willing to defend the proposition that "scientific quantitative generalization is not and cannot be broader or more exact than the mathematical statement." A little later in the paper Kelley continued:

I have quoted extensively from this early article and with a part of it I agree, - particularly with the last sentence: "The case is one of many where statistical ability, divorced from a scientific intimacy with the fundamental observations, leads nowhere." If Boring will interchange the words "statistical" and "scientific" I will still agree in this matter of quantitative scientific investigation."58

With this statement, Kelley turned to attack Boring's article, "The Logic of the Normal Law of Error in Mental Measurement."

Kelley stuck to his attack on Boring's deductive inclinations with respect to statistics. He set out to challenge the notion of "cogent reason" as Boring was using the phrase. But as we have seen in the previous chapter, Boring's position is not easy to understand. In part this is because Boring does not reject outright a frequency interpretation of probability, nor does he embrace a Baysian interpretation. Thus, Kelley is understandably confused by Boring's position and perceived it to be inherently contradictory.

Kelley began his criticisms of Boring's "Logic" paper by
first pointing out that in as much as statistics were inductive, Boring’s deductive understanding of them was simply too rigid and too narrow. He wrote:

[The] identification of frequency with probability necessarily enters into social statistics. It is the inductive element. It belongs in the process. . . . I know of no single practical problem . . . where earlier observed frequencies have not been taken as the guide to later expectation, and this is inherent in sound statistical procedure.59

Kelley saw statistics as inherently inferential. Indeed, he went so far as to say that using statistics in a deductive manner was to limit the procedures to one logical operation.

The scientist, Kelley thought, adjusts probability statements because the frequencies in the data (which have been repeatedly collected) support such an adjustment and not because of some apriori principle. For him frequency, in and of itself, told the scientist something about nature. By identifying frequency with probability, Kelley saw no gap between them. He wrote:

Even the Mendelians take an observed ratio of, let us say, 53:47 as suggestive of 50:50 and not of 75:25. The pure inductive treatment would be to take such a ratio as presumptive of 53:47, and to refine or correct the assumption or expectation as more evidence is gathered.

He commented further that the assumption between frequency and probability is never "indubitably established", as Boring recommended as proper scientific procedure. Rather,
Kelley thought that the "inexactness of earlier discovered frequencies" are questioned as indicative of "frequencies operating in general and for all time" only through "later experiments". These later experiments would then establish "a second and more accurate concept of the frequencies." In this way, and here Kelley again turns one of Boring's phrases on its head, "truth is gradually wrought out of ignorance."60

Boring reasoned that between frequency and inference there was a gap. He drew a distinction between a special kind of apriori theory and empirical, frequency theory. The apriori theory had to account for the latent traits or causal factors that determined the frequency distributions. He saw the process of identifying these causal factors as one that had to be different from the processes of describing the visible data. This former process established the cogency of expecting a particular frequency distribution in the data. Boring, as we have already seen, thought it was logically fallacious to hope to work back from frequency distribution to identify the underlying causal network.

Kelley regarded Boring's point as contradictory. Any attempt to link frequency to inference is necessarily based on uncertain hypotheses. He reasoned that Boring's effort to establish cogent reasons for expecting a particular distribution contrasted with the spirit of setting up a hypotheses. What he understood Boring to be saying was that "we have reasons for our guess." Further, he thought that Boring was attempting to establish certitudes, not
hypotheses, and this was a hopeless task and not at all what scientists do. Kelley wrote:

This position is untenable, as every act of life testifies. The bases of conduct are probabilities, not certainties, and the scientific procedure is to measure and express these probabilities, continually assuming the more and more probable as the basis for future expectation.

I'm not sure Boring would disagree that science must deal with uncertainties - probabilities - but he would be reluctant to identify sample frequency with population probability. At any rate, Kelley saw statistics as telling us something about the cogency of our hypotheses.

Kelley's next turn in his defense of statistics is revealing. He suggested that since we can not establish with any certainty which reasons would be cogent for accounting for a distributional form, we could still talk about "degrees of cogency and degrees of reliability of assumptions." He wrote that in "every case the probable error in the assumption increases as cogency decreases, and vice versa." He continued:

The real problem, then, is ever to decrease the violence of the assumption, and secondly, when either one of two assumptions will serve a given situation, to choose the more probable. Fortunately statistical method does not depend for its warrant upon such cogency as Boring advocates, and an assumption based upon the best obtainable evaluation of preceding experience is a sufficient starting-
point in the process of increasing knowledge at the expense of ignorance.

Again Kelley proceeds from a method - the calculation of a probable error - to make a statement about the validity of an assumption. Kelley never reasons outside of the collected data base.

Kelley's research program then was one that emphasized the collection and analysis of data. It was through data collection that new theories would derive their constitution, not through the reflection on apriori assumptions. Data preceded theory. The priority of the research process was to collect data, theorize later.

In the last section of his critique of Boring, Kelley wanted to address a point where he believed that Boring was openly contradictory. He also believed that this contradiction in Boring's position reflected his lack of understanding of statistics. Also, by ending on this note, he returned to the basic point he started his paper with: Boring is guilty of the very things he inveighs against.

At issue was Boring's comments in his "Logic" paper that suggested that it was not reasonable to use what today we call parametric statistics on rank ordered data. He thought that most data in psychology was rank-order and therefore it was inappropriate to use statistics that assumed interval data. Boring wrote:

What we must remember is that we are dealing with the statistics of medians, quartiles, contingencies, and correlation ratios; not with the statistics of averages, standard deviations, coefficients of
correlation and linear regressions. All those statistical constants, that imply a scale of equivalent units, violate in use the conditions of the case and lead to a precision of result that is an artifact.

Kelley’s response was simply to assert that much of what Boring was saying was wrong. He stated that even the interpretation of quartiles or medians required resort to assumptions of normality (or possibly to some other distributional form).

Kelley was wrong to suggest that the normal distribution was assumed in the derivation of Chi-square, at least not in the same way that it is in the derivation of a t-statistic or F-ratio. Rather, the normal distribution appears only as an approximation to the binomial distribution. Kelley believed in statistics and so he believed that the distribution of observed frequencies around an expected value would be normal across an infinite number of samplings. But this was really an empirical matter and not an assumption required for the derivation of Chi-square. Kelley’s interpretation of Chi-square points to his commitment to build into statistics a uniform, distributional order.

Kelley suggested that the normal distribution was implied in any statistical analysis. If one recommended the use of any statistics, even if they used it in a strictly descriptive manner, one was implicitly assuming that the normal distribution held for that data.

Kelley’s rejoinder was directed to demonstrate that
Boring's research prescriptives were not free of the very pitfalls he identified as being the property of statistical analyses. His point was not to target Boring as being an incompetent researcher. Rather he wanted to argue that psychological research had to proceed despite the messy entanglements of methodological assumptions. Whenever possible, Kelley pointed out that Boring was as committed as the psychologist-statisticians to the assumptions made in statistical analysis. It was only because of Boring's lack of statistical understanding, Kelley thought, that he believed that his own research practice avoided assumptions such as the normal distributions. Kelley ended the first part of his paper saying:

Our conclusion therefore is that none of the measures advocated by Boring is free of those things which he inveighs against. It is this situation in particular which leads me to suggest that Boring's criticisms were made without due appreciation of the foundations of statistics.61

Boring initially responded to Kelley's paper but this response was never published. Shortly after receiving Kelley's paper Boring wrote to him saying that he had to think over the criticisms "pretty thoroughly" before he could conclude whether his reply "would not be likely to be a misunderstanding of yourself." Boring thought that if he misunderstood Kelley, then Kelley would feel obliged to reply and this exchange might go on "ad infin." "You see," Boring continued, "I do not want to write thirty pages of quotations from you" only to have you reply with another
"thirty pages of quotation from me."

Boring certainly felt that Kelley misunderstood his point. He wrote to Terman that there seemed to be "only thirty-one things the matter with it" and that he would "probably find more later."62 By October of 1923 Boring had not yet formally replied to Kelley. He wrote to him a letter of explanation:

With respect to our "controversy" I perhaps owe it to you to explain why nothing has happened at this end. In the spring, when you were good enough to let me have your manuscript, I was sufficiently piqued by your detailed criticism to write a rejoinder. The manuscript when I finished was quite as long as your note, and I regard its length as a fatal objection. It would be silly to start a series of papers that does not approach zero as a limit.63

This was Boring's last correspondence with Kelley until 1926.
1. Sidney Pressey noted in "Empiricism versus Formalism in Work with Mental Tests," (Journal of Philosophy (1921) 18 393-398) that Boring's "The Logic of the Normal Law of Error in Mental Measurement" was "of the very greatest of theoretical importance." (p. 394, footnote 4.) Pressey argued for an empiricism in mental testing and to use statistics only when they aid prediction. The irony of Pressey's support for Boring's argument rest with the fact that he used it to support a system of mental testing that stood for everything Boring was opposed to. Pressey pushed for better prediction, not understanding of the underlying factors that made up mental test performance. In Chapter three of the thesis I outlined Ruml's criticism of Pressey.


3. G. Thomson to Boring, June 20, 1920. Harvard University Archives, Boring Correspondence. What Boring said in this final paragraph was very critical of the use of statistics. He ended the paragraph with two strong statements:

But, if in psychology we must deal--and it seems we must--with abilities, capacities, dispositions and tendencies, the nature of which we can not accurately define, then it is senseless to seek in the logical process of mathematical elaboration a
psychologically significant precision that was not present in the psychological setting of the problem. Just as ignorance will not breed knowledge, so inaccuracy of definition will never yield precision of result. (p. 33)


6. Ellis Michael responded to Boring with a note, Science (1920) 54 130.


8. I have gathered Kelley’s biographical material from various sources. Pertaining to his appointments at Texas, Teachers College and Stanford I have examined the letters sent to him by these institutions offering him these positions and the terms of employment. These letters are kept among Kelley’s papers at the Harvard University Archives. His appointment to the NRC committee on Mathematical Analysis of Statistics, I have drawn from a letter written to him by Augustus Trowbridge, February 1, 1921. In a letter he wrote to Helen Walker (November 30, 1926) he speaks of his own development as a statistician. I have also used some secondary sources: The Biographical Dictionary of American Educators edited by John F. Ohles (3
volumes), Westport: Greenwood Press, 1978 and the
International Encyclopedia of Social Science. New York: The
Macmillan Co. & The Free Press, 1968. Also a brief overview
of Kelley's life was provided in an obituary written by John

9. R.A. Fisher to T.L. Kelley, January 12, 1924. Kelley’s
Papers, Harvard University Archives.

10. K. Pearson to Kelley, March 17, 1924. Kelley Papers,
Harvard University Archives.

11. C. Burt to Kelley, January 14, 1944. Kelley Papers,
Harvard University Archives.

12. Lewis Terman, History of Psychology in Autobiography
(volume II) Worcester: Clark University Press, 1932, 297-
331, p. 320.

13. Terman to Kelley, January 14, 1920. Terman Papers,
Stanford University Archives.

14. Terman to Kelley, June 1, 1950. Lewis Terman Papers,
Stanford University Archives.

15. Terman to Boring, April 29, 1930. Terman Papers,
Stanford University Archives.

16. See Kelley’s notes for September 29-October 1, 1908 in
a notebook titled "Education I". Kelley’s papers, Harvard
University Archives.

Paper, Harvard University Archives.
18. Kelley to Helen Walker, November 20, 1926. op. cit. It is important to note that Kelley wrote this letter in response to a survey by Walker, who was collecting material for her 1929 book Studies in the History of Statistical Method. Kelley perhaps was interested in pointing out his connection with Pearson who was clearly a central figure in the history of statistics.

19. ibid.

20. Kelley to Dean Russell, January 15, 1920. Kelley Papers, Harvard University Library. The program was not endorsed.


24. Kelley’s approach is equally apparent as he related his research goals to Cubberley: "My idea was that I would spend my time here in developing an experimental and statistical method for determining the degree of uniqueness of different mental traits, with a view to securing a map or plan of mental life so that one could say function (a) is the most worth while trait of an individual to measure, function (b) the next most worth which, etc.—all with reference to its bearing upon scholastic, vocational and social success.” (Kelley to Cubberley, May 15 1922, Kelley Papers, Harvard
25. I will not deal in any detail with the split between Kelley and Spearman in this thesis as it involves differences over their calculations of probable errors and their different concepts of factor analysis. Since factor analysis as it was developed in America departed from Spearman’s approach, it requires a full discussion of the lines of departure. It is an important issue and I am presently working out these details for another project on the history of psychometric theory. This conflict, however, reflects that different approaches to the question of how statistical methodology informs theory—that is whether statistics are used as confirmatory or as a means of making inductions—is a point to which I will return.


27. E.P. Cubberley to Kelley, October 24, 1923. Kelley Papers, Harvard University Archives.


32. Ibid., p. 59. Here Ruml wrote, "When the results of several tests are combined, as for example, in the Binet or the Army Intelligence tests, the standing in the combination is again expressed in terms of a linear scale, not because we have analyzed our concept of and experiences with general intelligence and found it so expressible, but because our common methods of test measurement and combination preclude any other result."

33. Ibid., p. 60.

34. Kelley and Terman, "Dr. Ruml’s Criticisms of Mental Test Methods," op. cit., p. 461.

35. Ibid., p. 460.

36. It is interesting to note that even this example is chosen such that the correlation between mine ores would be heightened. That is, like the hip length, total height relationship, ores would occur in a mine in particular patterns, i.e., iron ore would appear with magnesium, tin with titanium, etc.. I'm not sure it was intentional on their part to choose examples that show different qualities that tended to be positively correlated, but it is clear that this type of metaphor fit with their sense of intelligence and the subscales that compose the overall test score.
37. Ruml referred to "derivative facts" in his paper, "The Need for an Examination of Certain Hypotheses in Mental Tests", op. cit., p. 57. He believed that forcing linear assumptions on the data misrepresented the data and created an understanding of intelligence (as a construct) which was misleading. See his letter to Kelley, Jan. 5, 1920. Kelley Papers, Harvard University Archives.


40. Ibid.


42. "Dr. Ruml's Criticism of Mental Test Methods", p. 461.


44. "Dr. Ruml's Criticism of Mental Test Methods," op. cit., p. 463.

45. This was in contrast to questioning, as Ruml was doing, whether or not these visible phenomena were anything other than artifacts - or "derivative facts", as he termed them.


48. Ibid., p. 411.

49. Ibid., p. 409.


53. Ibid., p. 337.


55. Ibid., p. 411.


59. Kelley, Ibid., p. 413.

60. Ibid., p. 413.

61. Ibid., p. 418.

62. Boring to Terman, March 26, 1923. Boring Correspondence, Harvard University Archives.

63. Boring to Kelley, October 6, 1923. Boring Correspondence, Harvard University Archives.
Boring did not respond to Kelley’s criticisms for three years. His delay, like Kelley’s, is difficult to explain. Clearly there were those who were interested in the issues being debated, but the debaters did not cooperate. It is as if they both felt that they made their points clearly and they simply could not put the issues in a better light. Neither were their respective research programs affected by the debate. But an opportunity to make amends, or to at least clarify their differences, emerged from another quarrel Boring was having with Carl Murchison.

What touched off their attempts at reconciliation was an exchange in 1926 between Boring and Carl Murchison which was published in the American Journal of Psychology.¹ This exchange was ostensibly concerned with the issue of the proper procedures for arriving at a scientific inference, though much more was involved in the debate.² Boring drew Kelley’s attention to his short "Note" by mentioning that he hoped his discussion of Murchison’s research might help resolve, or at least make clear, his disagreement with Kelley. Boring sent a copy of the "Note" to Kelley saying:

I cannot be sure whether there is still a point of difference between us or not. There ought to be, since I am not conscious of having taken anything back; but on the other hand I don’t see how there can be. Perhaps you will let me know by letter, unless you decide to annihilate me again in a further note. . . I apologize for lugging Murchison
in so vigorously, but it seemed too good a chance to accomplish an understanding with you.\textsuperscript{3}

Kelley responded to Boring's letter and stated that they still had a difference to resolve. The nature of this difference rested with their respective understandings of the probable error.

The probable error was a term coined by Bessel in 1815. In 1816 both Gauss and Bessel used the term, defined it and offered formulae for its calculation.\textsuperscript{4} The probable error referred to those deviations whose positive and negative values divided the normal curve such that one half of all the observations fell between these two values. Thus, the value of the probable error was some portion of what later became known as the standard deviation, whose positive and negative values encompassed 68.26% of the area around the mean in a normal distribution. More specifically, the value of the probable error was 0.6745 of a standard deviation. This is the simple statistical meaning of the word "probable error".

When used with astronomical observations the term "probable error" had a straightforward meaning. Galton protested its usage for anthropometric data, since he could not see how Nature could be in error just because all objects were not alike. He conceded to use the phrase only because it was "too firmly established to uproot."\textsuperscript{5}

In drawing out his differences with Boring over the interpretation of the probable error, Kelley proposed that they deal with a novel situation:

Suppose we measure a group of Zulus and a group of
Yaquis by means of a certain mental test. . . The Zulus average 2.0 higher than the Yaquis, and the formula for the probable error of the difference gives us a probable error of 1.0.6

He then took a statement from Boring's 1926 "Note" suggesting that any "evaluative departure from the statistical finding should always go in the direction of conservatism", and queried as to what was a conservative conclusion.7

When Boring used the phrase he was referring to the common situation (as he saw it) when a statistical test was carried out under conditions where the assumptions of such a test were violated. In such circumstances, Boring thought that the researcher should ignore the statistical finding and conclude that there were no group differences. This was the conservative decision. Boring recommended, in other words, that the researcher has to take account of all the "facts" - even the conditions under which a statistical test is conducted - and, at times, choose to ignore the results of a significance test. Thus, research conclusions were never arrived at only on the basis of a statistical summary.

"There is no apriori reason," Kelley argued, "why conservativism should indicate that Zulus and Yaquis should average the same, or that either one should be higher than the other." Kelley concluded that it was conservative to go with what the data indicated, keeping in mind "the mental reservation warranted by the size of the probable error."8

Kelley provided the following rationale for his position. He argued that if a sample was split in half,
similar results would be found in each half, thus:

From this kind of approach I believe that it is possible to make a conclusion about the relation of the particular group to the general group on internal evidence from the particular group alone.

Kelley ended his letter by stating that he was sure that if accessory data was available that he and Boring would "reason in a similar manner" but when it was not available, Kelley was "certainly prone to trust the given data for whatever it may be worth statistically."\(^9\)

"I surrender - not my point, but the hope of ever getting you to understand me." Such were the opening words of reply to Kelley's letter.\(^10\) Boring accepted the Zulu/Yaquis example and focused the discussion on scientific induction:

... if one is interested in averages one is almost sure (although not absolutely so) to be interested in induction. Will any crowd of Zulus on the average be better than any crowd of Yaquis?

Clearly the answer was no, for one could select the best Yaquis and compare them to the worst Zulus. "Well then," Boring went on after answering his rhetorical question, will any "fair sample" of Zulus be higher than any "fair sample" of Yaquis? Here your data do not define a fair sample; if you are as completely ignorant as I am you can only take what you happen to get; and I should not for a moment think of recommending separate classes for the Zulus and Yaquis, just sitting in Geneva and doing it on the
basis of this result.
Boring then suggested that he would "depart from the statistical finding in the direction of conservatism, and say that the result was not significant."

Boring clearly thought that his concerns were not easily addressed through a statistical method. What was required to decide if a sample was fair was the judgement of the scientist. Given no information on the nature of the samples, Boring regarded a "no difference" (conservative) judgement as appropriate. Nothing derived internally from the sample could speak to the point as to whether or not the sample was unbiased. Thus the probable error would be of little service in arriving at a decision as to whether or not the Zulus or the Yaquis were more intelligent.

Kelley’s argument was drawn strictly from statistical concerns and emerged out of a logic tied to split-half reliability assessments. His answer to Boring’s letter is even more revealing of their differences. "[Y]ou may be right", Kelley wrote, "for in your second paragraph you raise a point on which I certainly had not spoken."11

The point raised by Boring was that on a retest it was possible that the difference between the Zulus and the Yaquis would be reversed. Kelley proceeded:

I would very decidedly claim that the probable error of the mean is not the instrument to use in determining "the chance that in retesting the same crowds the difference would be reversed." The proper formula for this is:
Kelley pointed out that $r_{11}$ was the reliability coefficient of the test employed. He then added:

Reference to the formula shows that if the test employed is perfectly accurate, then retesting the same crowd would turn out in exactly the same way.

No, the ordinary formula for the probable error of the mean is the instrument whereby to determine the chance that in retesting a second crowd chosen in the same manner as was the first but ever so much larger, the difference will be reversed.

Kelley's formula and explanation are difficult to follow. This is due in part to the sketchy explanation he provided in his letter. Still, I think it is possible to get at what Kelley was attempting to say.

What Kelley was speaking to was the "probability" that the direction of the mean difference would be reversed if the same two groups were retested at a later date. He saw this not as a question about the variability of the groups per se, but this variability as it related to the reliability of the retest. Thus he calculated the probable error of the mean using the "standard error of measurement" in place of the standard deviation: instead of using $U_1$, he used $\sigma_1 \sqrt{1-r_{11}}$.

Some background in psychometric theory is required to understand Kelley's point. The observed score variance for a group is regarded as the sum of two other variance components: true score variance and error variance (the
variance around a true score due to measurement error). The "standard error of measurement" refers to this latter variance component, expressed as a standard deviation. The whole formula can be expressed as:

$$\sigma_1^2 = \sigma_\infty^2 + \sigma_{e1}^2,$$

where $\sigma_\infty^2$ = true score variance and $\sigma_{e1}^2$ = error of measurement.

What this formula tells us is that the observed score variance in a group reflects variability due to different abilities plus variability due to error. Thus around each true score, there was assumed to be an error distribution.

These concepts grow more complicated when we think of the true score of a group as opposed to the true score of an individual. The true score of a group was thought to be reflected by the mean of all the individual scores, plus error due to measurement. The true score of an individual was seen to be reflected in the mean of multiple testings, plus error due to measurement. Thus if the psychologist is working with a homogeneous group, all the individuals have the same "true score" and variability then is seen as measurement error. Thus in such situations, the probable error is used in much the same way as it was used by the astronomers. If the group is not homogeneous, then variability in this group reflects both variance due to true score and variance due to measurement error.

Returning to Kelley’s formula, we can see that he was talking about measurement error, $\sigma_1\sqrt{1-r_{11}}$. He must have assumed that Boring understood psychometric theory. I don’t think Boring was very sophisticated in psychometrics, and he
certainly missed the point. Kelley offered in his letter only one part of the equation for describing the variance components. What Kelley surely meant was:

\[
P_{\text{E-mean}} = \frac{.6745\sigma_1}{\sqrt{N}} = \frac{.6745(\sigma_\infty + \sigma_i \sqrt{1-r_{11}})}{\sqrt{N}}
\]

This formula is intended to be definitional rather than mathematical.\(^{15}\) With the formula Kelley offered in his letter, he referred only to the \(\sigma_1 \sqrt{1-r_{11}}\) — this was the probable error of measurement. If the sample was homogeneous (thus \(\sigma_\infty = 0\)), and the measurement was perfectly reliable \((r_{11} = 1)\), the formula above would reduce to zero. There would be no variability in the group — everybody would have the same score.

If the group was not homogeneous and the measurement (IQ test) was perfectly reliable, there would be no variability around the individual true scores that make up the group. Thus, the variance for the group would reflect true score variance. This can also be seen from the formula above, as the variance component due to measurement error would reduce to zero. This was not really the P.E., as the term was used by the astronomers, but reflected the groups variability.\(^{16}\) This appears to be what Kelley meant in his letter, though he did not lay out the details. Just relying on Kelley’s statements, it sounds as if he believed that there would be no variability. Clearly, there would still be group variability.\(^{17}\)

The idea that motivated Kelley to write the formula was that the probable error of the mean for a retested group must include information about the reliability of the test.
In his letter he only spoke of the case where the reliability of the test was perfect. It was clear to Kelley that if a test was perfectly reliable, the probable error within each group would remain the same on the retest; there would be no mean difference reversals.

The salient point for this discussion is that for Kelley statistical operations were sufficient in answering Boring. For him the probable error did account for the probability of new groups replicating the first finding. His new formula, which considered the reliability of the test, accounted for the probability of duplicating the original finding if the same groups were retested. It is as if Kelley did not understand the point Boring was making - that if your sample is biased, then so is your judgement of the groups if you rely only on the statistical finding of your sample.

In light of this, Boring's response to Kelley's P.E.\(\text{mean}\) formula is unexpected: "At last we agree upon one point, namely that the probable error of the mean is not an adequate predictive value. I give praise to heaven for so much progress."\(^1\)\(^8\) He then added that, "I take it also that there is little doubt that our faiths are not alike with respect to the more general problem of extending conclusions from samples to larger groups."\(^1\)\(^9\)

Clearly, Boring did not understand Kelley's formula. Of course Kelley's formula did not present the whole picture concerning the probable error either. That is, if the reliability of the test was equal to one, then Kelley's formula would produce a P.E.\(\text{mean}\) = 0. Clearly this is not
what Kelley meant to say. Rather he meant to say that the probable error of measurement would equal zero, and the variability of the groups would not be changed in a retest. Thus there was no chance of a mean difference reversal between Zulus and Yaquis on the retest. If the test reliability was less than one, then the standard error of the mean becomes larger, as the variance components – true score variance and measurement error variance – would contribute to the overall observed score variance. Thus, it is clear that Kelley did not abandon the probable error as a predictive tool, he merely modified its calculation to reflect the circumstances of the problem.

It is conceivable that Boring and Kelley did not fully understand the argument the other was making. To probe further into their different interpretations of the probable error requires an examination of more fundamental matters of measurement theory. Their respective interpretations of "variability" and the "unit of measurement" systematically tied their views of the probable error to their conceptions of applied science.

**Variability, Measurement and the Probable Error**

Principally the probable error was regarded as a description of the variability in a homogeneous sample. It provided the interval around the mean that accounted for 50% of the area under the normal curve. It had associated with it another dimension, one that was not entirely statistical. In astronomical data the probable error referred to errors of observation, with the implication that if another
scientist conducted the same observations with the same apparatus this scientist would produce data having a similar mean and probable error. Clearly, there lies here psychophysical assumptions about observation/judgement errors, assumptions which interested Fechner, Ernst Mach, James McKeen Cattell and F.M. Urban.

It was not these psychophysical assumptions however that interested most psychologist-statisticians. Whereas psychologists like Urban devised empirical tests to see if the normal distribution held for judgement errors on constant stimuli, mental testers were willing to assume that errors were distributed according to the normal law. For the mental testers the probable error designated the 50% region around the mean on a normal distribution. If the probable error was small, the implication was that measurement was accurate and the data were reliable. Thus the probable error came to be associated with both the accuracy of measurement and the reliability of that data. Indeed it was commonly felt in the mental testing community that if the study was done in the same way, with the same apparatus, on a homogeneous sample, a similar mean and probable error would result.

On this understanding, variability in the data was seen as being relatively stable. This was because it was believed that 'true scores' (a person's true or real intelligence) was constant and error variance (variability around a true score) was normally distributed. It was also assumed that the "probable error of measurement" for a test would be more or less constant across repeated measurements.
of these different samples. Thus the distribution for groups was assumed to be the same across samplings.

As to the question of what caused the variation around the mean (or true score, be it for an individual or for a homogeneous group) was debated among mental testers. That is, some regarded the variability as error, some regarded it as error plus some constant factors specific to a situation or individual.23 Regardless of these issues, most mental testers agreed that variability would be constant across samplings of the same population. Thus the probable error was seen as a stable characteristic of the data.

Although mental testers differed as to what factors may cause variability, it was seen as a rather aesthetic endeavour (of no practical consequence) to carry out detailed studies to isolate those factors. They conceived of variability as being the result of many causal factors, each having no more effect than any other. For example, H.E. Garrett in his book Statistics in Psychology and Education (1926) illustrates this point when he wrote:

... a man's height, or his weight, or the shape of his head, or his intelligence, or his eye color is determined, very probably, by a large number of factors which have approximately the same influence on the final result.

To this he added a note: "Should one or more of these factors have special weight the distribution will no longer be the probability type, but will be skewed or shifted over towards the upper or the lower end of the scale."24 If the distribution was normal, the attitude was that all of the
factors had an equal contribution. To disclose the underlying factors would be interesting, but the consequence of such a labor would have the mere result of removing the effect of a constant. Constants, being true to their title, would not be expected to fluctuate over samples. In this way, the probable error term gained its inferential power.

Returning to Kelley’s example, the discussion focused on the average score for the two groups on a mental test and addressed the question as to whether or not this data provided a sufficient basis for making a general statement about all Zulus and Yaquis. For Kelley the mean and probable error were inherently predictive instruments. Variability due to error had many causes and their effects balanced out in the long run. If the probable error was large, this did not mean that it was inherently less predictive, just less useful as a predictive instrument. Thus predicting an individual’s ‘true score’ on a mental test from the observed score (which was regarded as the true score plus error), would be difficult to specify under such circumstances. Still the probable error was interpreted to indicate, based on the sample data, what the mean score for another sample from the same population would be. Testers were practically minded and their interpretation of statistics reflected their pragmatism.

It was likely this type of confidence in the probable error term that led many researchers in the applied areas of psychology to eagerly calculate them. Applied researchers were exploring new methods of measurement, largely paper & pencil inventories, and required means through which these
methods could be accepted as both valid and scientific.\textsuperscript{25} By interpreting the probable error as a sign of both the reliability of the data and the accuracy of measurement, the step to applying such findings to a particular problem appeared to be more or less a logical extension. The calculation of probable errors became so popular by the mid-twenties, that L.L. Thurstone wrote:

Everybody is figuring probable errors. Statistical jobs in education are being justified as scientific, dignified, and trustworthy by the fact that probable errors have been figured.\textsuperscript{26}

Although Thurstone offered correctives to obvious abuses of the probable error, he still attached meanings that tied it to reliability concepts.

Boring differed fundamentally from the mental testers in his attitudes toward variability. Being an experimentalist he sought to control variability so as to attain an understanding of psychological phenomena. To control variability, one had to determine the factors that contributed to it and then limit the effects of these factors. He understood that causal variables determined the probabilities with which an effect variable would occur. Thus by understanding what contributed to variability was to understand something of the phenomena under investigation. To treat variability as remaining constant so as to enhance prediction was to allow practical matters to override scientific understanding.

A clear example of Boring's attitude toward variability can be seen in a letter addressed to Raymond Pearl, the
biometrician at Johns Hopkins. He was seeking advice on the nature of the psychometric function and related it to Pearl’s discussions of the biometric function:

The usual phrase of psychologists is that dispersion under a psychometric function is due to "the variability of protoplasm": a pretty phrase. It is plain as the nose on one’s face, however, that it includes variability of technique in the presentation of the stimulus and also variability in the report or response mechanism as well as the variability of protoplasm. Our hunch is to work the psychometric function under the usual conditions and then to introduce additional controls of stimulus on the one hand, all criteria of judgement on the other, one by one and see what happens to the dispersion.

It was through control of the situation, Boring surmised, that one could begin to understand factors that caused variability. He commented to Pearl: "I have a notion that most of the variability lies in these powers of technique, and that if they were controlled the dear old variable protoplasm might prove surprisingly stable."27

Boring was concerned that mental testers essentially ignored variability by assuming that it would remain stable for some future sampling. That is, by judging shifts in mean values with respect to probable errors, they assumed that these "significant" differences would hold for the future. Thus from sample data alone mental testers felt justified in making recommendations that they believed would
hold over the course of time. This was perplexing for Boring since he felt that it was through isolating what caused the form of a probability distribution that one gained grounds for making solid predictions. Boring commented to Lewis Terman that:

My fundamental faith is (as I have probably told you often before) that a detailed analytical study of the acts that make up an intelligence test by behavioral or introspective methods or both, would finally yield a much fuller knowledge of the nature of so called intelligence, and a knowledge that would throw it definitely into relation with the other body of knowledge that constitutes physiological psychology.28

As far as Boring was concerned intelligence tests in and of themselves revealed little about the underlying dimensions that may account for test performance. To analyze statistically such results held little promise of informing psychologists about the nature of intelligence or test performance. For Boring statements pertaining to the variability of test scores were of little scientific interest, unless one could account for the factors that produced the variability. Again he wrote to Raymond Pearl: "Seriously our mental testers are wasting a glorious amount of energy on figuring out results that don’t mean a damn thing when they are gotten."29

Both T.L. Kelley and E.G. Boring fashioned an interpretation of statistics that was suited to their ambitions. Boring saw statistics as descriptive. Inference
was a matter of trained scientific judgement conditioned by the data at hand; it was not a property of a statistical method. As we have seen in previous chapters, Boring regarded the laboratory as the institutional base of all science. Any movement away from the laboratory was considered by him as a moving away from scientific principles. In Boring’s interpretation of statistics - and this is clearly seen in his treatment of the probable error - his intention was to deprive them of their inferential power. In this way he provided further testimony to the indispensible nature of laboratory science.

Truman Kelley was a practical man interested in personnel management. It was of paramount importance to demonstrate that one could evaluate samples and make sound recommendations for future policy - this is the heart of applied science. For him statistics provided a tool, both elegant and efficient, for bridging the gulf between samples and populations.

Boring and Kelley’s different interpretations of statistics extends to their differences over just what should be considered a "unit" of measurement. Again we see that they styled their interpretations of measurement theory so as to promote the development of their research techniques as well as to maintain their institutional presence.

Units of Measurement

Measures of variability need not imply units of measurement, but often they do. Standard scores (difference
scores expressed in terms of standard deviation units) were growing in popularity during the 1920s among the mental testing community. This quest for a unit free science was emerging from the unsuccessful quest for appropriate units for mental measurement.

Hornstein (1987) argued that during the late nineteenth century, psychology redefined much of its content so as to make key properties quantifiable. That content which was not easily quantified was jettisoned from serious empirical study. She suggested further that most of this activity to quantitatively redefine psychology was initially centered in two areas: psychophysics and mental testing. Thus, the psychophysicists and the mental testers became the "arbitrators of what constituted appropriate forms of measurement for the discipline as a whole."30

Initially psychologists were concerned with the unit of measurement in a classical, physical science, sense. That is the unit was conceived of as a quantitative portion of that which was being measured. For example, in a foot there are 12 subdivisions which are inches. Or if something weighs 20 pounds, we can think of the basic unit as one pound. We could of course determine other units for these examples, but the unit is always a quantitative subdivision. On this view a total magnitude of the phenomenon being measured is a multiple or sub-multiple of the unit.

In psychophysics, Fechner conceived of the mental unit as a "just noticeable difference" (jnd) in the comparison of a stimulus with a standard stimulus.31 Thus any sensation intensity could be expressed as a sum of unit sensation
intensities, or a sum of jnds. The jnd was therefore understood as providing a subjective measurement unit. From these Fechner set out an equation for a psychometric function. He further assumed that the magnitude of change in the stimulus resulting in a jnd of sensation remained a constant portion of the value of the stimulus - \( M_{\text{JND}} \times \text{constant} \). He assumed this relation to be constant for any given sense.

Of course there were objections to Fechner's formulation. The most famous objection, and certainly the most cited, was issued by William James. He questioned that different sensations of intensity should be treated as a homogeneous group. Each sensation of intensity, James thought, was a qualitatively different experience. In the *Principles of Psychology* he wrote:

But really it has no meaning to talk about one judgment being bigger than another. And even if we leave out judgments and talk of sensations only, we have already found ourselves quite unable to read any clear meaning into the notion that they are masses of units combined. To introspection, our feeling of pink is surely not a portion of our feeling of scarlet; nor does the light of an electric arc seem to contain that of a tallow-candle in itself.32

Despite objections such as this, referred to in the literature of the day as the "quantity objection"33, psychologists worked with and developed the psychophysical methods of Fechner. Some psychologists like Titchener continued to try to determine a "true psychological unit". Boring cited with approval Titchener's attempt to develop the "sense distance" as a psychological unit in his "Logic"
paper. In light of new ideas as to the scientific function of measurement units, however, psychology began to move away from the classical concept of measurement.

Departures from classical measurement theory began to appear in the late 1910s and 1920s. These formulations were not clearly articulated though they appeared in the measurement literature of the 1920s. Perhaps the reason for this clumsy articulation rested with the persuasive power of the classical theory of measurement.

Traditional ideas of measurement held sway with respect to establishing psychology as a science within universities where the physical and natural sciences were better represented than the social sciences. Thus theories of measurement which competed with the classical models failed to gain much attention. This lack of attention impaired to some degree a clear articulation of these competing conceptions of measurement. Therefore, these new perspective on measurement were less successful at gaining access to the public forum; few listened, few argued, few abandoned traditional thoughts on measurement.

In terms of measurement theory, psychology was in trouble. The disclosure of "true" psychological units was appearing to be a hopeless task. Hornstein (1987) pointed out that the discussions of the "quantity objection" resulted in methodological procedures being treated separately from the theoretical discussions concerning measurement. Indeed, psychophysical methods developed in spite of serious theoretical objections about the meaningfulness of such operations. Thus methods were
gaining attention independent of the theoretical issues they were devised to address. Measurement theory was no exception. Methods of measurement were being developed, even though theory lagged behind.

Although I think Hornstein is correct in noting the split in psychology between the development of theory and methodology. It is equally important to realize that methodology was serving a slightly different function among some psychologists in the 1910s and '20s than it had in the previous half-century. As I have tried to show throughout this thesis, methods were being used to replace the psychologist’s judgement so as to render as more "objective" inferences from research findings. With respect to measurement, it was believed that units could be "objective" - and scientific - even if they did not fit with the classical notion of ‘unit of measurement’.

It was taken for granted by some (and these were growing in number) that if methods were managed properly, the interpretation of the meaning of the data would be self-evident. Such beliefs allowed for a reorganization of priorities in the hierarchy of scientific procedure. Matters of theory could now be placed below the collection of data since methods, applied to data, carried inferential power. Data collection would resolve theoretical dilemmas.

Not all experimentalists objected to how mental testers and some psychophysicists were restructuring research practices, but many - like Boring - did. These objections were not to the movement away from classical measurement theory. Clearly psychology had to adopt strategies with
regard to measurement that did not fit with classical perspectives. The meaning of a "unit" of measurement was being redefined by competing research groups.

There began to appear in the discourse of psychologists two ways of talking about measurement. Psychologists who tended toward a "manager of methods" approach to research talked about measurement in "operational" terms. That is, measurement was seen strictly in the terms of the operations used in the measuring process. Because measurement was restricted to the notion of assigning numbers to objects (in this case some psychological process or capacity), then—from an "operational" perspective—these number assignments had meaning only with respect to the measuring operations.

Those psychologists who were more aligned with an "objective arbitrator" model of research practice discussed measurement in "representational" terms. That is, number assignment had to represent the empirical, qualitative relationships that existed in reality. Their position was closer to the classical conception—because of its implied realism—but it did not conceive of the unit of measurement as a quantitative portion of the thing being measured.

The tension between operational and representational styles of measurement began to appear in the discourse of psychologists. But even this tension was an uneasy one as both sides shared the goal of developing alternative strategies to the classical perspective on measurement.35

In the next two sections representational and operational theory are further explicated. I place representational theory within the "objective arbitrator"
model of research practice, using E.G. Boring as a representative. In the section following this, I look at operational theory in the context of T.L. Kelley’s work as he represents the "manager of methods" approach to practicing science.

**E.G. Boring and Representational Measurement Theory**

Psychologists were aware of the problems associated with determining a "mental unit". Boring related to Urban that when he wrote his "Logic" article he was "very much worried" about the existence of true psychological units. He confessed that he was "biased by Titchener’s position in his Quantitative Manual" where Titchener sought to demonstrate the "scientific nature of psychology" by establishing "real mental units". But Boring noted that Titchener’s attempt, like all the others, "did not work out in practice."³⁶

Boring also told Urban that all his worries over this matter may have been "foolish ones." He wrote:

The popularization of the Einstein theory of relativity suddenly brings to light the fact that the physicist does not need to deal with the unit if he is to effect measurement.

Boring went on to suggest that measurement "consists simply in the establishment of coincidences." He provided an example that in a psychometric function we could determine the "coincidences between relative frequencies and values of stimulus." Any system of measurement then, Boring concluded, is "arbitrary and relative."

Boring stated that he reached this same conclusion in
his "Logic" article, but at that time he found it
discouraging. For him what was most discouraging was that
units were arbitrary. He felt that he could not avoid this
conclusion in light of the failure of psychologists to
establish a unit that could be properly called
psychological.37

Boring was moving in the direction of a representational
system of measurement. Accordingly he reasoned that
distributional forms were a function of arbitrary units of
measurement. At one point in his "Logic" article he wrote
that if one could begin with a "true psychological unit", a
distribution of mental capacity could determined from it.
This he thought was "the necessary scientific order."38 But
given the inability of researchers to find such a mental
unit, psychologists must use an arbitrarily chosen
"physical" unit. It was then the task of the researcher to
both determine the distributional form resulting from such a
unit and to establish just what these units meant
"psychologically".

Boring's objection to the mental testers was that they
assumed that the normal distribution held for their
arbitrarily chosen mental units - like mental age or the
speed with which a mental test is completed. He argued that
this was a question which was open to an empirical
demonstration and, he noted, the testing community had not
carried out such studies. Also Boring thought that mental
testers believed their units to be "measures" of mental
capacity. But these arbitrary units were not quantitative,
as far as he was concerned, at least not in the classical
sense.

Boring pointed out that there was no empirical evidence that indicated that an equal increment in these arbitrary units corresponded to an equal increment of the psychological entity being measured. Thus, as we saw in the last chapter, he objected to the use of statistical operations that assumed "equivalent units" on data where there was no evidence that such a condition existed in the data.

The manner in which Boring presented these issues suggests that he was developing a representational approach to measurement. He proposed that these "physical", "arbitrary" units allowed the psychologist to place the mental capacity on an arbitrary scale. All this provided was a rank ordering of performance. Thus Boring was suggesting that physical, arbitrarily chosen units can represent psychological quantities only on ordinal scales. Such units did not represent psychological quantity and to interpret them as such was logically fallacious.

Boring’s point was that the psychologist must determine and judge just how a unit is related to the mental capacity under investigation. That is, the units chosen should not be interpreted in such a way that they add quantitative meaning to the data; they should merely represent the data in different - perhaps more convenient - terms.

In Boring’s schema of measurement theory, if a unit was to be meaningfully employed it was required that it represented, in some way, the empirical qualitative relationships that existed in the "mental" data. Although I
think this point fades in and out of focus in his "Logic" paper, it was a driving force behind his 1921 paper, "The Stimulus-Error."41

In this paper Boring dealt mainly with problems in psychophysics because he felt that the effects of committing the stimulus-error were shown most clearly at this level. He believed that the "psychophysical experiment in its simplicity" represented the "ideal" in terms of its "control of conditions and adequacy of observation."42 He commented that the stimulus-error was apparent in both mental tests and in behaviorism in general. He added that an "extension of this discussion" of the stimulus-error to "higher processes" for the time being must wait.43 Nevertheless, he certainly believed that what he had to say about the stimulus-error at the psychophysical level applied to these studies of higher mental processes.

One of the definitions of the "stimulus error" offered by Boring was that it was a "meaning error". This consisted of "describing objects, reporting meanings" instead of "describing mental processes". He said that "[w]e commit the stimulus-error if we base our psychological reports upon objects rather than upon the mental material itself."44

What Boring was referring to was that to treat stimulus units as indicating something about mental processes led to errors of meaning, which clearly extended to errors in making scientific inferences. He was concerned that some psychologists ignored the stimulus-error in the belief that in so doing there was no real consequence in terms of the interpretation of their research results. These
psychologists, and Kelley was among them, were interested in the interrelations among response patterns which were measured in the units of the stimulus object. From the evaluation of these patterns of response, these psychologists were willing to say something about the subjects who generated the responses, but nothing in particular about what a specific unit represented to a particular individual.

Boring's strategy was to show that by not guarding against the stimulus-error, the psychologist risked collecting "response data" that was scientifically unreliable. Accordingly he cited various studies, some conducted in his laboratory at Clark University, that showed differences in psychophysical judgments as a result of controlling for the effects of subjects committing a stimulus-error. That is, when subjects were asked to ignore the stimulus and to instead concentrate on the mental sensations, they produced finer discriminations.

Boring reasoned from these studies that ignoring the effects of the stimulus-error led to results that rendered the "correlations between stimulus and response equivocal"; this in turn "jeopardized the rigor of conclusion that science demands." Boring suggested that in areas of fine discrimination if the experimenter did not guard against the stimulus-error, reversals occurred. For example, in determining a limen for a two point impression task, the subject would say under the same stimulus conditions that they sensed one point some of the time and two points at other times. The "inscrutable middle terms" - as Boring
called them - could be discriminated from one another better under conditions where subjects were asked to attend to sensations rather than to the stimulus. This led to correlations between stimulus and response that were less equivocal.

Boring argued that if the psychologist ignored the stimulus-error, the patterns found in the data of responses painted an inaccurate picture. The stimulus units did not represent accurately the mental processes. Boring concluded that the "danger of the stimulus-error" reduces to the danger that judgments of stimulus will prove to be "scientifically equivocal." Because these correlations were "equivocal", he surmised, they were "unscientific."46

To my knowledge, no one ever responded to Boring's "Stimulus-Error" paper. It is likely that it was read as a defense of introspection, which clearly it was. Since introspection as a psychological method was dying in the 1920s, the paper was largely ignored. Still the paper is of interest here as it sets out fairly clearly Boring's representational ideas on measurement. Basically he was arguing that if arbitrary units are to be used to make inferences about mental processes, then it is important that the psychologist determine the manner in which they represent mental processes.

S.S. Stevens, who was Boring's graduate student, went on to develop further the representational theory of measurement during the late 1930s. Stevens' (1946) paper, "On the Theory of Scales of Measurement", laid out the argument that measurement is basically the assignment of
numbers to objects according to a set of rules that allow the numbers to reflect the empirical relations between the objects. He carried his argument over to include restrictions on the type of statistical analyses that could be performed on data. He argued that the different scales of measurement (that is, different rules of number assignment to objects) placed restrictions on the type of statistical operations that could be employed.\textsuperscript{47}

Representational theory requires the psychologist to determine the empirical relations in the data so as to enable the researcher to prescribe a number assignment that does not distort the data. In this approach to measurement the researcher remains as an important determinant of the meaning of the data. Statistical operations, as well as number assignments to objects, serve a descriptive function. All of this is in keeping with Boring’s position that maintained the priority of the researcher (rather than methods) in drawing inferences from data.

T.L. Kelley’s Operationalism

I have drawn out these points on Boring’s drift toward a representational system of measurement in order to contrast it with Kelley’s drift toward an operationalism. An operational system of measurement maintains that the interpretive priority be given to methods. I think this is in keeping with Kelley’s approach to research which in turn, was typical of much of the work being done in the applied research areas.

It is important to keep in mind that the term
"operationalism" was not available to Kelley, nor had Bridgeman articulated this perspective with respect to measurement theory. Still, the ways in which Kelley discussed measurement, the ways he contrasted his views with Boring's ideas, suggest that he was moving towards an operational theory of measurement.

In his response to Boring's "Logic" paper, Kelley discussed the problem of the "unit" in psychological measurement. He wrote that although it "might seem axiomatic that there can not be a science of quantitative measurement until ... there is established a particular unit of measurement," a close examination of the issues did not require this. He then stated his position: "The existence of the science does not lie in the units employed, but in the relationships which are established as following after the choice of the units." 48

It is apparent from Kelley's remarks that, like Boring, he could not see holding back the establishment of a science of psychology until a "true" mental unit was found. Both Kelley and Boring, and certainly most psychologists, had accepted that for the present psychology had to work with arbitrary units. But this move to accepting arbitrary units was treated differently by different groups of psychologists.

As I argued previously, for E. G. Boring arbitrary units did not mean that psychologists could neglect their psychological meaning, even if these units could not be considered as quantitative in the classical sense. That is, these arbitrary units had to be placed in the laboratory so
that their effects and meanings within a research context could be determined. For Kelley, arbitrary meant something quite different.

Kelley noted that in mental measurement, "starting with units however defined," psychologists could still "establish important relationships between phenomena measured in these units." He suggested that the choice of a unit of measurement should be determined solely on the grounds of "utility." He wrote:

> Without denying the possibility of other workable systems, it seems that, in a civilization such as ours, steeped in the elementary associative and commutative principles of arithmetic and algebra, much is to be gained in simplicity and accuracy of interpretation if the units employed in mental measurement obey these well-known laws.

Kelley emphasized the importance of mathematical compatibility in one's choice of units. At times he reflected an indebtedness to the classical conception of measurement and therefore, short passages which inquire about the quantitative meaning of the unit appear in his work. He proposed, however, that such inquiries should follow after a unit has been found which provides a "picture of mental relationships" which is both "simple to comprehend and to treat statistically."

For Kelley the measurement unit determined the mental relationships, in a strictly operational sense. That is, if the use of a unit resulted in a pattern of responses that discriminated among subjects, these patterns could be
studied. If relationships could be drawn from these patterns, and if these relationships were (in Kelley’s words) of "value" in "doing the work of the world", then such a unit served the development of a quantitative psychology. In a lecture titled "In What Units Shall We Measure Intelligence and Achievement" Kelley (1928) stated:

My discussion therefore starts with measuring devices that are valuable, and it does not need to start with any hypothesis that we know just what the valuable thing is that we are measuring.

In this same lecture he remarked that "the measuring device as a measure of something that it is desirable to measure comes first, and what it is a measure of comes second."52

For Kelley, a quantitative psychology studied those relationships that result from using a particular unit. If the patterns that are disclosed using such a unit prove to be "valuable", hold up over time, and are reliable, then these data provide the basis for establishing a quantitative psychology of mental measurement. The actual meaning of the unit was not of principal importance until the patterns of relationship were established.

Kelley used statistics to study these relationships in the data. He used them often in an exploratory fashion. That is, sometimes he discovered in the response patterns correlations that he would not have predicted. This reinforced his commitment to using statistics as the primary scientific tool in generating theories. If a particular unit led to useful patterns of correlation in the responses from subjects, the unit served a scientific purpose. What
would happen if another unit was used in its place, or what another unit meant under the same conditions, was not of interest. What was of interest was whether or not a chosen unit enabled psychologists to make better predictions or allowed them to provide better guidance. These conditions were judged primarily by looking at correlation coefficients and probable errors.

The assumptions statistics made about the data did not worry Kelley either. Units could be chosen, he thought, that reflected these assumptions. If a meaningful response pattern was not produced using such units, some other unit could be chosen. Thus an interesting difference between the representational approach and the operational approach is apparent. In the former it is presumed that numbers maintain the relationships in the data; in the latter, numbers are assigned to maintain the solvency of statistical criteria.

A good example of this latter procedure can be found in Kelley’s work. He was concerned with developing an intelligence scale that would gauge the influence of nurture on test performance. In outlining the procedures for such a test he reasoned that the "correlation between sibs as given at birth would not be expected to change throughout the life of the couple, if allowance is made for growth." From this he extended his argument:

The measure of this correlation will be weakened by any improper units of measurement. If therefore we start with units as given by a test, and alter them so that a higher correlation between sibs is
obtained, we shall be altering them in the direction of natural units of native ability in the function tested.

Using this approach Kelley determined that the "Stanford Binet" was a test "largely influenced by native ability and, to a lesser extent, by nurture." He carried out a study to enhance the Stanford Binet's sensitivity to native ability by altering the "Stanford Binet units" so that they would provide a maximum correlation between sibs.53

The value of a measurement unit was judged against the correlations that resulted from the response data. If the correlation patterns that were produced by using a particular unit held across time and resulted in small probable errors, they were judged as valuable units of measurement. The meaning of a measurement unit was not judged according to how well it represented a mental process, but by its ability to produce correlational patterns that aided clinical and practical prediction. The "true" psychological meaning of a unit was something that for Kelley and other mental testers could wait while the "work of the world" marched on.

The difference between Kelley and Boring with respect to the unit of measurement can be further highlighted through Kelley's discussion of Boring's "Logic" paper. Kelley noted that Boring seemed to be confused when he discussed the unit of measurement. That is, some of the time Boring argued that the unit used in psychological research was necessarily arbitrary. At other times in the article, Kelley noted that Boring reverted to placing restrictions on the use of a unit
as if it were not arbitrary. What stirred the discussion was Boring's comment that if a psychometric function follows the normal distribution then one has to know the stimulus unit that is being used. This was because different units were differentially represented in mental processes. Thus if it was determined that judgments of "heavier" and "lighter" followed the normal distribution around a standard for gram weights in a psychophysical experiment, this may not hold for kilogram weights. As Boring saw it, grams are represented differently in sensation than are kilograms. Thus, if we use a unit, we have to determine empirically the psychometric function and from this the psychologist can gain some sense of how this unit represents a mental process.

Kelley objected that as long as the units used were linearly related, he could not see a problem with changing units. Thus if a psychometric function held for grams, then it should hold for kilograms as it is linearly related to grams. As Kelley so boldly stated: "... any linear transformation of the scale does not change the form of distribution." This is certainly true mathematically. As far as psychometric functions go, Boring argued, this was not the case. Just because units are mathematically related in a linear fashion did not imply that such units were so represented in mental space. Thus just because grams render a phi-function of gamma did not mean that kilograms or pounds would result in phi-functions as well. Thus to proceed to apply linear transformations on data without first determining their psychometric function, was to place...
mathematical elegance ahead of psychological meaning. Indeed Boring reasoned that to draw inferences from such data transformations led to inaccurate conclusions. He regarded such research practices as reflecting a trend to "a-psychologize" psychology.55

Boring was not being inconsistent, nor was he restricting the "very great freedom" psychologists had in choosing a unit. Kelley perceived him as being inconsistent because Kelley reasoned about units in a more operational fashion. He perceived units first in relation to mathematical formulae, and secondly in relation to mental processes. Boring did just the opposite. He perceived the meaning of units in terms of their representation of mental processes and then, given these restrictions, in terms of their mathematical fit.

Conclusion: Operationalism, Representationaism & Units

The direction psychologists began to move with respect to measurement theory, and the question of the "unit" in particular, was toward "representational" theory and "operationalism." Although neither of these was very well developed within psychology during the 1920s, by the 1930s these conceptions dominated psychologists' ideas about the meaning of measurement.56

Representational theory suggested the view that measurement was an operation that assigned numbers to objects but did so in such a way as to preserve the qualitative relationships among the objects. That is, number assignment should not add arithmetic properties to
the objects being measured. Rather, number assignment was carried out through a system of rules that was thought to reflect (or represent) the empirical relationships already present in the data.

Operationalism, on the other hand, was more concerned with the procedures through which objects were assigned numbers, and less with what the numbers represented. The numbers were not meaningful except in light of the operations. The strong thesis of operationalism suggested that objects - or their properties - did not have an existence outside of the operations through which they were made visible.

The tension between these two conceptions of measurement is apparent. Representational theory suggests that the empirical qualitative relationships among objects are somehow knowable and when they are made known, numbers can be assigned as a shorthand to describe them. A fundamental error would occur if the objects were assigned numbers in such a way that they suggested arithmetic properties that were not represented in these objects. Conclusions based on such number assignments would be misleading.

On the other hand, operational theory suggests that it is through methodological procedures that objects are placed into relationship with each other. We cannot and do not know how they are related outside of the operations that relate them. Thus if numbers are assigned in a consistent manner, and if one is true to a pure operationalism, one cannot suggest that properties have been added to the data. Operationalism is sort of a 'behaviorism of methods'.
Summary

I have attempted to show in this chapter that interpretations of variability and units of measurement are intertwined with the social relations and interests of competing groups of psychologists. The probable error term was seen as a predictive instrument by the mental testing community because of the ways in which they interpreted variability. That is they assumed that variability was constant across different samples drawn from a population. They reasoned that this stability was owing to the many causes which acted on a population. Furthermore, these causes were thought to be equal in their effect and therefore rendered normal distributions in the response scores of their subjects. If distributions deviated from normality, mental testers considered that this was due to one causal factor having a greater effect that the others.

Given such an understanding of data, statistical constants like the probable error suggested what might be expected from a future sample under the same conditions. If the probable error was small, then the measurement was regarded as precise and predictions to the future would be more accurate. Data collection and statistical analysis began to be emphasized in research practice.

Methods gained inferential power. More and more the concern in psychology departments focused on a fuller training for students in the use of statistical methodology. Statistics were promoted in such a way that they were seen as inherently inferential. But their usefulness became even
more apparent in addressing the problems psychologists were facing in measurement theory.

The quantitative unit of measurement in psychology was much debated throughout the nineteenth century. The problem was in defining just what constituted a mental unit. Fechner proposed the "just noticeable difference" as such a unit, but the critics raised what became known as the "quantity objection". Titchener attempted through his conception of the "sense distance" to create a unit that was "mental". But this too failed in practice. Nevertheless, Psychologists believed that even though they did not have a true mental unit, they were doing science. They had to re-think the meaning of measurement.

In the drift away from classical conceptions, two styles of measurement became pronounced in psychological research. The first was representational theory and the other operationalism. In some respects the two attitudes to measurement shared similar goals. They were, after all, designed as alternatives to classical theory. They both differed from classical theory in that they regarded the unit of measurement as arbitrary.

This concordance between operational and representational styles of measurement did not constitute a territory of firm agreement. From a practical perspective, operational and representational theory interpreted differently what "arbitrary units" meant within a research program.

Boring articulated a representational attitude toward measurement. Although he agreed that units were chosen
arbitrarily, their psychological meaning had to be empirically determined. Thus units served as the starting point from which the researcher could experimentally determine the psychological meaning of the units chosen. This determination of the psychological meaning of a chosen unit in turn placed restrictions on how the unit could be analyzed.

For Boring then, not all arbitrarily chosen units could be analyzed using statistics that required equal intervals. Indeed, Boring believed that most measurement in psychology represented a rank ordering. He reasoned that the type of analysis used had to reflect that the units represented an ordering of mental capacity by rank and not by an exacting magnitude. Thus most statistical analysis in psychology, he thought, could not draw on the calculation of means, standard deviations, probable errors, and coefficients of correlations. These statistics required equal-interval data and, Boring argued, arbitrary units can not be assumed to represent mental capacity or mental processes in this manner.

From Boring's perspective scientific generalization was a bigger question that mathematical description. The psychologist had to determine what the arbitrary units meant in psychological terms. The qualitative differences in mental processes had to be disclosed in order to determine what number assignments through units of measurement meant in psychological terms. The psychologist could not just treat units in an arithmetic capacity as this led to a scientific precision that was an "artifact". Kelley and
the mental testing community in general did not think of the unit of measurement in representational terms. Rather they took patterns among measurements as indicative of underlying mental process and capacities. Kelley reasoned that given that the unit of measurement was necessarily arbitrary, the researcher should always choose units that work well in mathematical formulae. In this way the relationships that are made visible through subjects' responses could be analyzed using statistical methods such as correlation coefficients.

Stephen Jay Gould is critical of the testing community. He commented on the tendency for mental testers to find the meaning of measurement in the pattern of correlations and not in the measurement units themselves. He wrote:

The idea that we have detected something "underlying" the externalities of a large set of correlation coefficients, something perhaps more real than the superficial measurements themselves, can be intoxicating.57

This tendency stems from the awkward kind of operationalism that was emerging in the 1920s in psychology. What made it awkward was that on the one hand numbers were assigned to objects so as to meet the assumptions of statistical analysis. Thus the arbitrary units were not directly interpretable with respect to the objects they were assigned to. However, the correlational patterns, being on a broader plane of analysis, were seen as interpretable and therefore meaningful to theory construction. Thus statistics began to be interpreted more as measurements themselves and less as
statements about measurements (since on the unit level
interpretation of measurement was clouded).

This operationalism where the units were evaluated
against their fit with mathematical criteria maintains the
priority of methodology. That is, Kelley’s understanding of
measurement reflects the "managers of methods" approach to
research. The inferences are simply given in the data and
revealed in an orderly fashion through statistical analysis.

The tension that emerged between the research traditions
represented by Boring and Kelley spilled over into the
politics of everyday activities. The final chapter treats
three specific incidents which demonstrate the day to day
effects of methodological differences that are supposed by
many to be on the plane of abstract ideas alone.
Notes


2. I have discussed this debate in a paper presented to the Canadian Psychological Association, Toronto, 1986--"The Politics of Statistical Inference: A Case Study of a Debate between E.G. Boring and Carl Murchison." A more detailed account of the debate appears in the following chapter.


5. Ibid., p. 53.


8. Ibid.

9. Ibid.

10. Boring to Kelley, May 18, 1926.


13. 'True score' is a term used by psychologists to represent the hypothetical conditions of measurement of a ability, attitude, or trait without error. That is if an individual has a true score of 100 for his IQ, he may or may not score exactly 100 on an IQ test, but usually some score there about. Thus there is measurement error around the true score. True scores are hypothetical, unobservable and generally estimated on the basis of observed scores. True score is similar to an "expected value" in a statistical sense.

14. This point may be difficult to grasp. It is similar to Galton’s concept of variability. Stigler’s discussion of Galton, and in particular his reprint of Galton’s "quincunx" and his explanatory notes, nicely illustrates the type of thinking that was reflected in the mental testers. See S. Stigler, *The History of Statistics: The Measurement of Uncertainty Before 1900*, Cambridge: The Belknap Press of Harvard University Press, pp. 265-299. The point is that although there is the sense that variability reflects an array of different talents (in this case, different IQ’s), variability also reflects error of assessment of these talents.
15. I composed this formula to make the point clear. It did not appear in the literature at the time, though it is true to what mental testers were saying in situations where measurement was being made on heterogeneous groups. The formula Kelley reported in his letter hold for a homogeneous sample. I call the formula definitional as I’m not sure about the mathematics involved in its derivation, though I think it make definitional sense.

16. Here the mental testers were following the lead of Galton. For a good discussion of the Galton’s perspective on variability and "error" as it differed from Quetelet’s perspective, see Victor Hilts, "Statistics and Social Science," in R.N. Giere and R.S. Westfall (editors), Foundations of the Scientific Method: The Nineteenth Century, Bloomington: University of Indiana Press, 1973, 206-233. A problem with reading some of the early work of the mental testers is that their presentation of statistical arguments are sometimes vague. This is understandable since psychometric theory was in the early stages of development and they were working through their ideas concerning the meanings of 'true score', 'true score variance' and 'error variance'.

17. This whole matter of error in psychometric theory is has a history of confusion with respect to the correct use of the proper term. Kelley debated the matter in 1923 with Holzinger, when the latter proposed a new formula for the probable error. See Karl J. Holzinger, "An Analysis of the Errors in Mental Measurement," Journal of Educational
Psychology (1923) 14 278-288 and Kelley's reply, "Note Upon Holzinger's Formula for the Probable Error," Journal of Educational Psychology, (1923) 14 376-377. This set off a series of exchanges. The conflicts over the proper error term continues to this day, see Frank J. Dudek, "The continuing Misinterpretation of the Standard Error of Measurement," Psychological Bulletin (1979) 86 335-337.

18. Boring to Kelley, June 2, 1926.


20. The continental direction in statistics was, in the main, depicted by the study of the homogeneity and stability in repeated trials. This was in contrast with the British tradition which proceeded by making precise prior distribution assumptions—usually of normality—and detailing the consequences. The attempt to minimize the effects of making apriori assumptions was a achieved through the study of "robustness". For a cursory discussion see C.C. Heyde & E. Sencta, I.J. Bienamme: Statistical Theory Anticipated. New York: Springer-Verlag, 1977, in particular pp 49-57.

21. F.M. Urban (1910), "The Method of Constant Stimuli and its Generalization." Psychological Review, 17, 229-259. Urban was interested in errors of observation and their psychometric distribution. He wrote, "The judgments of a subject who compares two stimuli under well-defined and constant conditions, have the formal and material character
of those chance events which are spoken of in the calculus of probabilities." He tied his findings to Gauss and astronomical observation and suggested that such psychometric observations were of consequence to the physical sciences. Cattell felt that psychophysics, though not necessarily Fechner's elaborations, had important implications for the physical sciences. He wrote, "it is within the province of psychology to supply physics with the formulae it requires in eliminating errors of observation in special cases." ("On Errors of Observation," American Journal of Psychology 5 285-293; quoted in Michael Sokal's biography of Cattell, chapter 4, "Establishing the New Psychology at the University of Pennsylvania." [1889-1891], to be published in Spring of 1987.)

22. This notion is not logically tenable since the apparatus, or the manner in which a study is conduced, may place constraints on just what data is produced. That is, the methodological procedures of a study may introduce systematic errors so as to restrict the range of values that can be expressed by the data. Therefore, although replication is achieved, we have gained little understanding of the phenomenon under investigation.

23. A good example of such a discussion can be seen in Kelley, "Note on the Reliability of a Test: A Reply to Dr. Crum's Criticism," Journal of Educational Psychology, (1924) 15 193-204.

25. K. Danzinger has drawn attention to the fact that it was in the applied areas that statistics flourished. With the development of applied psychology the research interest shifted from the individual to group data. See K. Danzinger, "Educational Administration and a Critical Shift in Psychological Research Practice," presented to the Cheiron meeting, Vassar College, June, 1984.


27. Boring to Pearl, Feb. 10, 1922.

28. Boring to Terman, May 26, 1926. At this time Boring just published a paper with his student Helen Peak with the title: "The Factor of Speed in Intelligence." Journal of Experimental Psychology 9 71-94.


31. A jnd refers to the smallest difference in value between two stimuli that leads to the discriminability of the stimuli. Fechner conceived of all jnds as being equal in terms of their subjective magnitude. For further discussion of jnd see Howard C. Warren, Dictionary of Psychology, Boston: Houghton
Mifflin Company, 1934.


35. I suggest this because often in Boring and Kelley's exchange of letters and within their articles are statements that suggest that their opponent would agree under some specified set of circumstances. I realize that some of this discourse is likely due to of the context of the textual material - politeness in debate - but I think it reaches deeper. They sensed their differences were great, but could not always reason their way through their differences as there was much in what the other was saying that they agreed with. Still there were profound differences that stood out, even to them.


37. See his statements in particular on p. 31 of "The Logic of the Normal Law in Mental Measurement," op. cit.

38. Ibid., p. 30.
39. It is clear from Boring’s passage in his "Logic" paper (p. 33) that he was referring to ordinal scales, though the term "ordinal" was not in usage until S.S. Stevens used the term in the late 1930s.

40. In a way, Boring resorts to an argument similar in form to Wundt’s psychophysiological parallelism. That is a representational approach to measurement asks that the physical unit not be identified with the psychological event. In psychophysiological parallelism it is maintained that physical events merely occur in correspondence with mental events, they are not mental events. Thus just as it is a mistake in representational theory to confuse numbers with mental quantity, it is a mistake to confuse a physiological analysis as being psychological.


42. Ibid., p. 451.

43. Ibid., see his discussion on pp. 451 & 460-462.

44. Ibid., p. 451.

45. Ibid., Boring discusses these studies in some detail between pages 462 & 470. The most convincing studies for him were the ones that investigated the determination of limen in a dual impression task. That is the subject had to determine when two simultaneous impressions were perceived as one. The manipulation was to decrease the distance between the two points until they were perceived as being
one point.

46. Ibid., p. 470, 471.


49. Ibid., p. 418.

50. Ibid., p. 419.

51. Ibid., p. 432.

52. T.L. Kelley, "In What Units Shall We Measure Intelligence and Achievement," in Scientific Method: Its Function in Research and in Education, Columbus: Ohio State University Press, 1929, 84-112. This is a published volume of Kelley's lectures.

53. Kelley, "In What Units Shall We Measure Intelligence and Achievement," op. cit., p. 104. He claimed by adjusting what he called the "Stanford Binet units" he increased the correlations between sibs from .62 to .72.

54. See Kelley, ibid., pp. 415-417.


56. There has been a recent resurgence of the debate over the relationship between scales of measurement and statistical operations. Principally it has been Gatio and Townsend who have openly debated the points in the journals.

Joel Mitchell, "Measurement Scales and Statistics: A Clash of Paradigms," Psychological Bulletin (1986) 100 398-407 has reviewed these recent debates and has suggested that no mention is made of the "classical" theory of measurement. His paper testifies to the fact that operational and representational theories of measurement dominate the modern literature and I would suggest they have done so certainly since the 1940s. He commented: "Though now largely supplanted in the minds of psychologists by the representational and operational theories, vestiges of the classical theory still persist in psychology." (p. 404). I have found Michell's paper to be quite formative on some of the issues I have discussed in this section.

CHAPTER 6

Professional Politics and Methodological Debates

The debates I have examined in the previous chapters focus on the tensions that emerged out of the split between experimental and statistical approaches to psychological research. Basically the quarrells centered on questions of how to generate knowledge that would serve in the building of an applied psychological science. But with different conceptions of applied research, which drew on overlapping notions of science and what was meant by "expert knowledge", the debates did not always effect a clear division even among the participants. On close inspection of these debates, neither this methodological divide nor the differences over the place of the researcher in applying knowledge, received unambiguous expression.

The reason for this is that methodology issues and professional politics interacted. The participants in the methodological debates were being tactical as well as strategic. What occurred in the discipline as a result of the split between experimental and statistical psychologists was a display of the messy business of week-to-week professional politicking.

Psychologists were not just interested in the "right ideas" they were interested in promoting their own careers as well. Thus disputes over methodology were not just epistemological; they had a direct bearing on careers and on who could best produce psychological knowledge. The two approaches had different implications for how psychologists should be trained as well as where they should be trained.
The two research perspectives competed for funding, space and—to some degree—institutional presence.

In this chapter I look at these methodological differences from the perspective of the week-to-week politicking that went on in the discipline. My treatment, however, is somewhat one-sided. That is, I examine the methodological debates largely through the political activities of Edwin G. Boring.

The reason for this bias is due to one important fact: Boring left a record of his activities. He was the secretary of the APA during the early twenties and he kept many letters from this period, some of which were quite explicit in terms of their politics. Other experimentalists were also involved in such activities, but I have had neither the resources nor the opportunity to read through their correspondence.

In the discussions that follow, Boring provides a focal point. His critics respond and accuse, he defends and attacks. Although the generality of this history is limited, it does suggest that politics were important to these statistical debates. Boring, after all, was the main critic of statistical methodology in the 1920s.

E. G. Boring and Applied Psychology

John O’Donnell has argued convincingly that applied psychology served as a support system for many psychologists. For instance, John B. Watson wrote to J.H. Hollander complaining that he could make more money writing articles for popular magazines than teaching courses:
I guess it's up to me to do my bit, but when I think that Harpers have been after me for some articles, and that I can write an article and get $250 for it, it does seem as though something were wrong with university renumeration.¹

Joseph Jastrow, in his "Autobiography", referred to applied research as the "pay vein that supports the mine".² Robert Yerkes, an experimental psychologist interested in animal research, worked at Boston Psychopathic Hospital because it was pointed out to him that such a move would be beneficial to his career.³

This move into applied research left the laboratory somewhat barren of personnel and under funded. In 1924 Boring and Lashley discussed the merits of holding a conference where problems in experimental psychology could be discussed. Lashley wrote saying that he needed a group "from which I could get some real constructive criticism." He also thought that the group should be restricted to "men of research experience" and should exclude "purely applied research."⁴ Boring's reply reflected the predicament of experimental psychology:

In general I agree with the tenor of your remarks, but I doubt if we could make it as highly selected as you suggest. . . The fundamental difficulty now-a-days seems to me to be that there are not very many psychologists deeply in research, and there are only a few when you have subtracted the technologists.⁵

For Boring such a group would serve several purposes.
The first was that it would allow experimentalists to get better acquainted. It would be a show piece for experimental research which would not only gain more attention for such research but would stimulate junior colleagues. Also, Boring noted, the "laboratory atmosphere" would be exhibited as well as its role in breaking down "school prejudices". The implication was that experimental research could unify psychology.

Boring realized that experimental research required more financial and institutional support. America, however, was a difficult place to gain assistance for endeavors that appeared to be far removed from the practical. Certainly this was Boring's perspective. He wrote to Gerardus Heymans (Professor of Philosophy and Psychology at the University of Groningen) that it was common for Europeans to think that America was wealthy and, it followed from this, that such wealth could be channeled to support all sorts of scientific projects. He pointed out that this simply was not the case. "It is a wealthy country", Boring wrote, "but it is not a country that appreciates science except in its applied phases and it is sometimes quite impossible to get even small sums for scientific purposes."

Boring reiterated this point to his friend Karl de Schweinitz:

We need more both of apparatus and for assistants if we are really to go on running a laboratory and not be driven out of it into our offices to write books. The writing of books without contact with research results is the sort of bosh that Munsterberg turned
Boring appealed to de Schweinitz as he was someone who knew how to approach the wealthy. "We need if anything worthwhile is to be done," he remarked, "to find some convincing way of putting the ideals of pure research before possible donors."

To emphasize the importance of laboratory research to de Schweinitz, Boring turned the discussion to a consideration of the relationship between pure research and applied knowledge. He wrote that he had "no prejudice against practical psychology" but "we all think in very much larger terms." What he meant by this was that "the consistent building up of the structure of scientific psychology" invariably led to offshoots of practical value. He noted that Binet, who he claimed was a "pure psychologist", just "happened" to develop intelligence tests "out of his work in pure psychology." Boring ended the letter with a suggestion that he would like to study the "nature of intelligence" in the laboratory. Karl de Schweinitz replied to Boring's petition saying that he would "spread the gospel of research for the sake of acquiring facts."

Boring did more than appeal to those who had influence with a financial elite. He cooperated with other experimentalists to secure for strategic value the secretariat of the APA. As O'Donnell has pointed out, Langfeld appointed Boring as his successor to the position, saying: "I should much regret seeing things in the hands of an applied man." It was a matter of maintaining research standards. Boring passed the position on to John Anderson,
an animal experimentalist. In 1925 the post went to Boring's friend and Tichener's former student, Sammuel Fernberger.

O'Donnell (1979) has pointed out that the experimentalists used their position on the Executive Council of the American Psychological Association to enhance the status of research. In 1921 this council prescribed that membership into the APA be conditional on published research and the attainment of the Ph.D. degree. In 1924 all but one member of the Council was an affiliate of Tichener's society of experimentalists.

Boring also maneuvered to have an experimentalist nominated for APA president. Although he was unsuccessful in getting Titchener the nomination, he did try to persuade Bentley--Titchener's former student--to run for the presidency. He was disappointed when Bentley withdrew. He wrote to Knight Dunlap:

I pass on to you in confidence what Bentley has written to me in confidence. I have to do this because I have already told you that we have an invitation to Illinois next year. Bentley now writes that he feels extremely impatient of the development of tests and also that the internal situation at Stanford discomforts him. He does not believe that he has any chance of being elected president, and he does not want to be host to Terman.12

Terman was elected APA president.

In the remaining part of this chapter, I argue that
Boring’s politics were a part of his criticisms of the use of statistics in research. That he chose to critique statistics was not incidental to his concerns over the status of experimental psychology. Statistics were being used as a basis for making scientific generalizations. The laboratory, as far as he was concerned, was the location from which causes could be isolated and generalizations sanctioned. To shun the tradition of laboratory research to gain appeal (and funds) for practical research, was no less than taking short cuts to the formation of scientific inferences. He scolded those who were party to such actions and he interfered with their careers.

Boring attempted to discredit mental testers as they were considered by him to be promoters of "the lustiest form of mental measurement."¹³ In what follows I examine three incidents: Boring’s review of Carl Brigham’s book on race and intelligence, his attempt to prevent Ben Wood, a former student of T.L. Kelley, from being admitted to the APA, and his attempts to discredit the research of Carl Murchison and thereby, the psychology department at Clark University. The first two I can deal with only very briefly as there is little documentation available. The third instance portrays a debate between E.G. Boring and Carl Murchison over statistical method. It provides a good example of the politics involved in their methodological debate.

Carl Brigham and the Politics of IQ

After World War I, Brigham broke down the results of the Army mental tests (Alpha and Beta) to find out if there was
a pattern in IQ scores that would correspond to racial makeup. He published his results in a book titled *A Study of American Intelligence* (1923). He emphasized that the decline in America’s intelligence was owing to the influx of low IQ immigrants. In 1930, in light of recent findings on combining test scores, he recanted. He wrote:

> This review has summarized some of the more recent test findings which show that comparative studies of various national and racial groups may not be made with existing tests, and which show, in particular, that one of the most pretentious of these comparative racial studies—the writer’s own—was without foundation.14

It was Truman Kelley’s books *Interpretation of Educational Measurement* (1927) and *Cross Roads in the Mind of Man* (1928) that provoked Brigham’s retraction, not the comments of the mental testing critics.

Brigham’s book was received with enthusiasm. Stephen Jay Gould (1981) suggested that Brigham’s book was the "primary vehicle for translating the army results on group differences into social action." Gould contended that it "materially affected the establishment of national quotas" on immigrants by providing what was taken be a "scientific backing" for eugenic arguments.15 Daniel J. Kevles (1985) also noted the impact of Brigham’s book. He saw it as feeding the "eugenically minded" public just what they wanted to hear and thereby gaining an impact beyond the quality of the arguments presented. He commented:

> High scientific authority — geneticists,
psychologists, anthropologists - drew upon expert 'evidence,' notably Henry Goddard's I.Q. test of immigrants and Carl Brigham's analysis of the Army intelligence test results, to proclaim that a large proportion of immigrants bordered on or fell into the "feebleminded" category and that their continued entrance into the country made, in Robert Yerkes's phrase, for the "menace of race deterioration."\textsuperscript{16}

Nevertheless, there were critics from various areas of psychology. F.N. Freeman and W.C. Bagley, both leading educationists, criticized the book.\textsuperscript{17} Freeman thought Brigham read the data so that it would support his theory about racial intelligence. Bagley made a similar criticism, noting that Blacks in the north who were literate had higher IQ scores than literate southern Whites. Such a finding he thought lent support to the effect education could have on intelligence test performance.

Boring also wrote a criticism of Brigham. The tone of this particular criticism was entirely unexpected. Boring agreed to write a positive review of \textit{A Study of American Intelligence}, but within a brief period of time (three months) he changed his mind. This transition is difficult to account for, from a number of perspectives. On the one hand, Boring was a critic of mental measurement methods and it is surprising that he would agree to write a positive review of Brigham's book. On the other hand, having agreed to promote the scientific merits of the book, Boring panned it in his review.\textsuperscript{18}

Yerkes had asked Boring to review Brigham's book. He
expected a supportive gesture on Boring’s part. There was reason to believe that the review would be positive. On receiving Brigham’s book, Boring wrote to him:

The American Intelligence is here and I looked at it through and read the last chapters with my wife last night. I think I was as critically minded as I was last spring, but I am disposed towards nothing but the heartiest congratulations. You have presented a difficult subject in a wonderfully clear and careful manner, and if our colleagues McDougall and even Terman would follow your methods, there would be less for Lippmann and Dewey to jump on. Congratulations then!

Boring went on to point out that he had an "arrangement" with the New Republic whereby he could ask to review books. After talking with Yerkes, Boring told Brigham that they "agreed" that writing a review of the book "would be an excellent thing to do." He added that from the perspective of the New Republic it would be a timely review and "I would like to put your book forward as a type of thing that a scientific psychologist can do."19

Brigham wrote back that he thought his book was as "airtight" as he could make it, "but not as airtight as it should be." Nevertheless he thought that he proved "that alpha and beta measures intelligence, or at least something that makes for educational and industrial success." He thought that such traits were "desirable" in immigrants and that he had "proved" that "our recent immigrants have less of this than the older immigrants or the old stock."20
There was nothing to suggest that Boring would write a negative review. Yet, he did just that. Even his title-"Facts and Fancies of Immigration" - was derisive. In the opening paragraph Boring commented that "the thoughtful reader" would likely "refuse to follow" Brigham. The reason for this, according to Boring, was that Brigham did not reason logically in drawing universal statements from particulars. This was the book's only "important error". He suggested that the "plain" facts that Brigham kept referring to in his book were not always so plain:

They are plain when applied to the army recruits; they are exceedingly obscure when applied to a description and to a prediction of American civilization.21

Boring was troubled by Brigham's reliance on a statistical analysis of the data. He noted that Brigham found statistically significant differences in test performance among the immigrant groups. Brigham found that the "Nordics" were superior to the "Alpines" who were in turn slightly better than the "Mediterranians". Boring commented: "The differences are statistically significant, but do the results mean anything as to the nations from which we should draw our immigrants?"22

Boring thought that Brigham failed to address this question in a convincing manner. This "failing" was not "Brigham's fault" according to Boring. Rather the "trouble" accrued from the data. Boring commented that the army results "were not collected with an ultimate scientific analysis in mind". It was a mistake, Boring thought, to
attempt to produce scientific generalizations from data that was not collected with a scientific purpose in mind. He concluded his review saying:

We need ever so much more information and, especially, data collected under better conditions. There is a mountain of statistical material in the army report. That in this case the mountain could bring forth only a timid mouse may be due to the fact that mountains for all their size do not necessarily have leviathans in them.  

Boring warned Brigham about the review and sent him the manuscript. It was in March when Boring wrote to Brigham saying that he could not write a positive review; his criticisms, he added, were not to be taken as personal and he sent his apologies. He continued:

It is the problem of the validity of a sample when conditions of sampling are not safely known. I believe emphatically in statistics as a tool for description, but it is a very dangerous device for the extension of knowledge from observed particulars to unobserved universals in which we are interested.  

This entire matter between Boring and Brigham is confusing. Boring knew that the army data was problematic from his scientific orientation. Also, he had claimed to have read the concluding chapters of the book and yet he still agreed to write a fairly positive review. What happened between January and March? Did Boring plan his switch over? Had he not read the book when he agreed to
review it and, upon reading it, came to the realization that he could not support such research? If this was the case, what would lead him in the first place to suppose that Brigham would be careful in his conclusions?

The answer to any or all of these questions is problematic. After Brigham published his book, Terman was so impressed that he offered him an associate professorship in applied psychology at Stanford.25 It was in this connection that Terman wrote to Boring asking what he thought of Brigham. Boring gave him a sound recommendation. He then added:

I say all this in spite of a conviction that Brigham’s book on American Intelligence is youthful and the conclusions unwarranted. It misses in my opinion (Yerkes to the contrary, notwithstanding) being a good scientific job. Brigham must have been anxious to get positive conclusions of considerable social importance, and hence sacrificed his logic to this end.

This last point caused Boring to pause over his recommendation:

I cannot decide whether this is a fault of youth, or whether it is a definite deficiency in Brigham. He is so practical and administratively efficient, that it would be no surprise to find him lacking in scientific brilliance.26

Terman was aware of Boring’s predilections concerning mental testers and would likely disregard these latter comments. After all, Terman had written to Yerkes that he
thought Brigham's work was "excellent and convincing." Yet Boring's comments do help us in understanding a little more about Boring's review. He thought that the danger of applied research lay in its lazy attitude toward facts. That is, if the results expected are obtained, and if such results are of practical value, the analysis stops there.

Applied psychology, according to Boring, had to be as tough minded about facts as laboratory psychology. Mental testing was like any other topic in psychology in that it was available for a more rigorous analysis. Boring thought that psychologists should get the mental tests into the laboratory and begin the long difficult task of isolating the factors that determined test results. In 1926 he managed to get one of his graduate students to carry out a study in the lab on intelligence test performance. He wrote to Terman:

My fundamental faith is (as I have probably told you often before) that a detailed analytical study of the acts that make up an intelligence test by behavioral or introspective methods or both, would finally yield a much fuller knowledge of the nature of so called intelligence, and a knowledge that would throw it definitely into relation with the other body of knowledge that constitutes physiological psychology as one finds it in the ordinary textbook.

With this letter Boring included the article he wrote with Helen Peak (1926) titled, "The Factor of Speed in Intelligence."
It is possible that Boring believed that Brigham would write about intelligence test data in what Boring would have considered to be a responsible manner. After all, Brigham was a student at Princeton where Howard Warren, a member of Tichener's group of experimentalists, directed the laboratory. Warren was competent and would engender an adequate scientific apprenticeship program for a young psychologist. Certainly Brigham's early work indicated a commitment to the laboratory.

Brigham had written an article and a monograph on mental tests in 1914 and 1917 respectively. The first was titled "An Experimental Critique of the Binet-Simon Scale," and the monograph (which ran 254 pages) was titled "Two Studies in Mental Tests." Both were critical of intelligence testing, though not without optimism. They represented what Boring certainly would have considered to be a discriminating encouragement for those involved in the testing field. With respect to the Monograph, Brigham thought that testing provided a reasonable diagnostic tool, but urged that caution in using the tests was necessary. He wrote:

The concept of "mental age" was exceedingly easy of comprehension, no apparatus was needed, and the scale has now become the common property of all. This development or overdevelopment has taken place inspite of the warnings of the authors themselves [Binet and Simon] and the psychological fraternity in general.

He added to this that the question as to whether or not the
Binet scale provides an "accurate measure of intelligence," could be decided "only by the study of the individual tests and the factors underlying them." The interpretation that Brigham meant that these factors had to be isolated in systematic laboratory studies would be consistent with his training under Warren. He concluded his monograph with the following statement:

Inferences from the nature of the tests to the nature of intelligence are of course uncertain, for we know very little about the mental processes involved in the tests. The mere fact that a psychologist classifies a test as involving a certain process does not prove that that process is involved.

Based on such a treatment of mental tests, it is possible that Boring assumed that Brigham would be circumspect with the Army Alpha & Beta test data.

It is possible that Boring, in light of his conversations with Yerkes, was satisfied that A Study of American Intelligence would be conservative with respect to the promotion of the results of the Army mental test data. Brigham was not conservative in this respect. But surely this would have been apparent in the final chapters which Boring claimed to have read.

It has been pointed out that psychologists at this time generally agreed on the directions indicated by Brigham's book. Russell Marks pointed out that many psychologists, for instance the "environmentalist" Bagley, drew the same "social implications" as Brigham. Psychologists differed
in terms of their proposed etiology of such problems not their sociology. Immigrants were perceived to be a social problem, and many psychologists accepted that immigration restrictive policies must be enacted.

It is possible that Boring agreed with Brigham on many matters concerning immigration. When he wrote his first letter to Brigham perhaps he reflected this enthusiasm of agreement, while overlooking the basis on which Brigham drew his conclusions. Given Brigham’s training and previous publications, Boring may have assumed that he would be more circumspect in his data analysis. It is possible that after reading the book, Boring realized his mistake. But there is another matter which I think colors the story.

Boring wrote to Brigham in January of 1923. In February of the same year, Boring received a manuscript from Truman Lee Kelley criticizing his papers which inveighed against the use of statistics. After three years without a response from the mental testing community, Kelley was now informing Boring that this silence had reached an end. Kelley’s manuscript, which was only a few pages shorter than Boring’s lengthy 1920 article, "The Logic of the Normal Law of Error in Mental Measurement", challenged Boring’s argument at every turn. He opened the paper with the comment: "Boring’s conclusions are generally destructive, and tend to leave one with the feeling that there is no sound statistical basis for mental measurement, and little for other psychological measurement."34

Boring was about to embark on a public disagreement over methodology. Until this point, it is possible that Boring
believed that his case against statistics was fairly sound. After all, Kelley had promised to respond earlier but never delivered. Kelley—a respected psychologist-statistician who had just published a major textbook in this field—was now prepared to challenge Boring publicly. He wrote:

I am enclosing the long promised or threatened criticism of your article in the *American Journal of Psychology*, together with an elaboration on what I call a constructive program in mental measurement. I have sent this article to Professor Titchener for publication in the *American Journal of Psychology*. Should you have any comments, either for print, or otherwise upon it, I should be very glad to learn of them.

Boring was not prepared at this time to address Kelley in detail, though he eventually wrote a rejoinder. This, however, was never published. Boring did have before him a document, Brigham’s book, which could serve as a point to remind readers of his previous convictions.

Boring responded to Kelley a week later saying that he doubted if their differences could be settled in print. The inconsistencies which Kelley saw in Boring’s arguments were not inconsistencies to Boring. "You see," he wrote, "I do not want to write thirty pages of quotations from you in reply to thirty quotations from me and have the thing go on indefinitely." He thought that it would be a waste of journal space. Boring went on to mention "a couple of things" that had a "bearing on the basis of our misunderstanding."
The first of these was a concern about the "unit in mental measurement." The other issue was the "problem of sampling and generalizability." In discussing this second point Boring mentioned that he was engaged in writing a review of Brigham's book. He told Kelley that he agreed to write the review because he had the notion that he "would be favorably disposed to most of his conclusions." But he was disappointed:

The more I read the [more] negatively inclined I become, and it is all because he makes the logical slip that I tried to distinguish between when I contrasted "scientific" with "mathematical" significance. He is interested in dealing with classes of which he has only samples and the relation of the sample to the class cannot be mathematically determined because no data are available.

Boring added that Brigham was "worse than this because he has samples of samples of samples of the total group with which he has to deal."35

In the April 25, 1923 edition of The New Republic Boring published his unanticipated review of Brigham's book. It is reasonable that Boring thought that Brigham would have written a more cautious treatise and that he was disappointed that Brigham had fallen into a statisticians way of reasoning about data. It is also reasonable to suggest that Boring had second thoughts about writing a positive review of A Study of American Intelligence after he received notice that his criticisms of statistics were going
to be fully addressed by T.L. Kelley. Boring may have acted to get something in print to remind psychologists of the pitfalls of making inferences from statistics in and of themselves. When Boring wrote to Brigham telling him of his change of mind, he added that the problem of making inferences was a "real issue" and that Kelley was publishing a "long article criticizing me for my stand on the issue." He added however, that he was "perfectly convinced" that Kelley was "wrong."36

There is one obvious irony in this story. Boring's criticisms of statistics, and Brigham in particular, did not move Brigham to retract his findings. Kelley's work, which was critical of Spearman, caused Brigham to reflect and later to recant. It was the fight within a research perspective that had the greatest effect, not the criticisms from the outside. Perhaps the reason for this is that within a shared perspective there is agreement as to what is considered evidence as well as what is considered to be a challenge to that evidence. The irony is extended even further when one considers that it may have been Kelley's criticism of Boring that led to Boring's uncharitable review of Brigham.

Nevertheless, it is clear that Boring acted to place Brigham's work in an unfavorable light. Just what lay behind Boring's writing of his review of Brigham is a collage of politics that is difficult to sort out. One thing is apparent: Boring sought to damage Brigham and the types of methods used in mental measurement circles.

Another interesting aspect of this story is that when
Kelley first wrote to Boring informing him of his critique, he included a comment concerning his former student Ben Wood. Kelley was under the impression that "there was something back of" the APA Council's decision to reject Wood's application for membership.37

Ben Wood, Boring and the APA

In 1920 Terman wrote to Kelley asking for the name of a "young man" for a position at Stanford.38 Kelley recommended W.J. Osborn and Ben Wood. He thought that Wood was the better choice: "I consider him a very high grade young man--more capable than Osborn." The only reservation Kelley had concerning Wood was that he "might get into trouble because of his outspokenness."39

Kelley thought Terman should take the risk and hire Wood. While an assistant professor at Texas, Kelley employed Wood as his assistant and thought highly of him as both a teacher and a researcher. Together they drew up the "Kelley-Wood Statistical Tables." At the time of Terman's request, Wood was working as Thorndike's assistant and was employed as a psychologist at a private school in the New York area. Wood earned his Ph.D. under Thorndike at Teacher's College.

In 1923 Wood was assistant professor of Collegiate Educational Research in Columbia College.40 His book, Measurement in Higher Education was published in this same year. Terman, the editor of the series devoted to "measurement and adjustment", wrote a special chapter to be included in Wood's book. He claimed that he did this
because Wood requested it and because of his own "conviction that the significance of the contribution of this book to higher education" was "so great as to justify its introduction to the American Educational Public in a more attention-compelling manner."41

Despite Wood’s record, and the kind of support Kelley and Terman were willing to place in him, he was denied membership to the APA. On February 20, 1923 Kelley wrote to Boring in response to the Council’s decision to hold over for further consideration Wood’s application for membership. He wrote that he "was a little surprised by the position taken," since it seemed to be "a very punctilious sort of position in the first place." "In the second place," Kelley went on to write, "I am certain that a similar position has not been taken with reference to other recent nominees."42

Boring replied with a lengthy explanation as to why Wood was not elected to APA membership. Primary among the reasons was that Wood had not yet met Columbia’s requirement of publishing his dissertation. Technically, he had not received his Ph.D. He met all the other formal requirements. Boring suggested that it was likely that Wood would be accepted as a member next year, but for now he had to be excluded. He remarked that the publications Wood submitted were substandard. Boring denied anything "being back of" Woods denial of membership.43

Before writing to Boring, Kelley discussed the matter with Terman. Terman wrote to Kelley that he found it "hard to understand" the Council’s thinking that Wood "was probably trying to put something over on them." He
suggested to Kelley that there was nothing that could be done about their decision "until there is another opportunity to present his name to the Council next Christmas." Terman added that he told Boring what he thought of their action.44

The entire incident is sketchy; there is not enough information available to get at what was going on. Certainly Wood was premature in his application to the APA, but only in a technical sense. His Ph.D. research, which formed the basis of his book *Measurement in Higher Education*, was judged by both Terman and Kelley to be of a high standard. Wood had also worked with Kelley and published a respected set of statistical tables. Yet, inspite of this, the APA council deferred on Wood's application for membership.

Of course formally the Council was correct; but Kelley's notion that it was a "punctilious" action was likely to be correct as well. Kelley's accusation that there was something behind this decision may not have been far off the mark either. As O'Donnell (1979) pointed out, the experimentalists attempted to control the APA council during these years and set restrictive policies that were employed in full force to control applied psychologists.45 Boring's attempt to prevent Murchison's membership to the APA is a clearer case of his interference. This particular case is interesting in that Boring explicitly used a methodological critique to forestall Murchison's acceptance into the APA.

Clark University, Carl Murchison and E.G. Boring

281
In 1926 E.G. Boring and Carl Murchison debated the merits of differing methodological approaches to psychological research. Although their debate received little attention, it provides an incident which demonstrates the complexity of placing methodological debates in their context. The debate centered on the split between experimental and statistical approaches to research. But as we have seen in the cases of Brigham and Wood, this divide in methodological perspective is meshed with the social relations of political opportunitism.

In the case of Boring and Murchison quite local institutional politics interact with a longlasting (perhaps still important) divide as to the sites where valid psychological knowledge could be produced, as well as the types of competence required to certify it. I argue that these points were not merely abstract matters of epistemology but had a direct bearing on careers and on who could best produce psychological knowledge.

The Debate

In 1926 Boring published a note in the *American Journal of Psychology* directing criticisms at the research of Carl Murchison and his associates. At issue was the appropriate means through which scientists were entitled to make generalizations. Boring set out the case:

Generalization is important because there is not science without it. . . . Scientific induction is, however, precarious because there is not for it the same carefully standardized technique that there is
for description. Consequently, nearly every induction starts life as a hypothesis, and becomes a law only if it survives the perpetual process of verification long enough to arrive at maturity.47 Boring’s criticism of Murchison was that he short circuited the inductive process and arrived at unwarranted conclusions.

Murchison responded to Boring by suggesting that he was not competent to make the accusations voiced in his note. He wrote:

"It is a serious matter for a man in one field of psychology to assume to teach a man in another field of psychology how to play his game. Such helpfulness is meritorious and should be appreciated if it is valid and can be trusted . . . can I trust Boring?"48

His answer to the question was of course no. He called on the opinion of T.L. Kelley for his comments on Boring’s methodological perspective. Murchison quoted Kelley from an earlier critique of Boring’s views and isolated a couple of statements that, withdrawn from their context, were quite critical:

In discussing a systematic statement by Boring, Kelley says: "The procedure described is such as I believe a competent scientist never resorts to . . . ‘Absurd’ hardly characterizes this procedure."49

Murchison objected to Boring’s criticisms by saying that the critic himself drew conclusions in his own research from methodological procedures that were suspect.
At issue between Boring and Murchison was a procedural disagreement as to how a researcher should move from the characteristics of a sample to the characteristics of a population. For Boring inference could never properly be statistical only. For Murchison, inference was indeed a statistical matter and deficiencies in existing methods of inference could be remedied by more sophisticated technical tools.

Boring's Critique: Description vs. Prediction

As I have indicated throughout this study, Boring was an ally to traditional research methods. He was also the most outspoken critic of the use of statistics in psychological research, a task he performed intermittently from 1919 to 1960. His 1926 note, which drew Murchison out for criticism, was also a comment directed at psychologist-statisticians.

For Boring, statistics were useful descriptive tools. As such, statistics did not constitute a full analysis of the meaning inherent in the data. To use statistics as a predictive tool was seen by him as akin to a leap of faith. Thus Boring reasoned that the inductive use of statistics was like arriving at conclusions based on nescience, not knowledge of the data. The probable error of the mean was a statistic that was often used in a predictive manner. In his 1926 Note he described the probable error of the mean as the statistician's "sleight of hand". He commented further that in his earlier papers he argued that statistics such as the probable error were treated as "measures of 'significance'" but were "actually only descriptive accounts
of the difference in relation to its dispersion and to the number of cases involved." He added that since they were descriptions, "they are prior to the inductive process and their determination is not an inductive process." 51

"It seems plain to me," he continued, "that we can not possibly make a conclusion about the relation of the particular group to the general group on internal evidence from the particular group alone." 52 He capped his argument by suggesting that the probable error of the mean could not serve as a predictor of the variability of the means of new groups. He stated his rationale:

... the central tendencies of new groups depend, not on anything inherent in the first group, but on the changed conditions that make the groups new and not identical with the old. 53

One would expect, given Boring's argument, that he would highlight it by citing psychologists who had misapplied the probable error of the mean, or would draw attention to those who clearly reached beyond the warrants of their data due to an over confidence in the inductive nature of statistics. Instead Boring used a couple of studies by Murchison as examples to be set up for ridicule.

If Murchison and his graduate student coauthors had used the probable error in presenting their data, one could understand why Boring would single these studies out. But they did not calculate probable errors! They calculated percentages and made crude comparisons to normative proportions from a larger population. Indeed Murchison & Gilbert perceived their method to be a first step in
identifying causal factors. They wrote:

If in any crime group the Law of Probability alone determines the relative proportion of single and married men, that proportion should be 60 to 40. If the probability is not sole cause determining proportions, then the proportion of single to married will be greater or less than 60 to 40.⁵⁴

No doubt statements such as this would trouble Boring as it was a clear demonstration of the misuse of the "Law of Probability". What is curious about Boring’s 1926 Note was that he did not select out such statements for criticism. Paradoxically, though not inconsistent with his position, he criticized Murchison for not providing a complete statistical description of the data. He pointed out that Murchison should have calculated the probable errors of the mean before making comparisons between groups.

Boring was correct in pointing out that comparing groups without the benefit of a measure of the variability was bound to result in unwarranted conclusions. Yet such a criticism from someone who had frequently cast aspersions on the probable error is a little unexpected.

The disquieting thing about Boring’s critique of Murchison was its "damned if you do, damned if you don’t" nature. He had Murchison in two ways. He criticized him for not calculating the probable errors. If Murchison had done so, Boring would have still criticized him for using the probable error of the mean differences as a prescriptive statistic. Indeed, after calculating the probable errors for Murchison’s data, Boring went on to argue that
inferences based on such statistics alone would be wrong. "My point is," he wrote,
that I do not think that the effectiveness of statistics for proving the generality of a
difference should increase in proportion to one's ignorance of the special conditions that both
determine and limit a particular sample!55

Murchison's Answer to Boring

In Murchison's reply he accused Boring of not following his own strictures pertaining to scientific generalization. The strategy Murchison employed was to argue that mental testing was as scientific as laboratory methods. Furthermore, he suggested that Boring's comments would invalidate the conclusions he [Boring] arrived at in his own laboratory investigations. Basically, Murchison argued that the scientific strictures for making inductions which Boring insisted on were unrealistic and represented a program that no scientist would adopt in practice.

Murchison statistical point of view was apparent in his criticisms. He began by noting that all of Boring's recent publications arrived at general conclusions based on the information obtained from only a few subjects. He wrote that, "after his investigation of appetite, based only on three subjects, Boring makes a universal conclusion . . ."56 Likewise, Murchison criticized Boring's study on "Introspection in Dementia Precox," noting that a general conclusion—that the precox could indicate a general train of consciousness—was arrived at on the basis of the responses of only "eight" subjects.57 Finally Murchison
suggested that Boring—in his paper "Sensations of the Alimentary Canal"—arrived at "sixty universal generalizations!" on information collected from only seven subjects.\(^5^8\)

Murchison’s criticisms focused on the number of subjects and not on the detailed methodological procedures Boring used in these studies. For example Boring’s article, "The Sensations of the Alimentary Canal," revealed within its fifty-seven pages many details pertaining to the methods used.\(^5^9\) The study elaborated on the statements made by the subjects involved in the study, of whom Boring was one. Yet Murchison does not mention this proliferation of methodological detail. Rather, the focus was on the number of subjects in the study.

In an effort to compare his approach to Boring’s, Murchison noted that his studies were based on a sample of seven thousand criminals. For Murchison large sample sizes *ipso facto* allowed one to draw tentative conclusions from the data. The implication was that Boring’s conclusions were unwarranted given the small sample sizes used. What Murchison reflected was an acceptance of certain statistical/probabilistic thinking that was promoted by Quetelet and had been picked up by psychologists.

For Quetelet large samples indicated the "constant causes" that were present in a population. Information based on large samples reflected information on *types*, not individuals.\(^6^0\) Being that ‘*types*’ are generalizations, and given that it was accepted by statisticians that increasing sample size functioned to cancel out individual variability,
it is possible to understand Murchison's underlying rationale for his criticism of Boring's small sample sizes. Also it is equally possible, in light of this, that Murchison did not understand Boring's criticisms. Specifically he may have passed over Boring's comments that data derived from a sample pertains only to that sample and not to populations--unless there has been some attempt to control variability and that the entire process is open to replication. Murchison may have failed to grasp Boring's point since he accepted that if a sample is large enough, it would reflect general types and not individuals.

Summary of the Debate

Murchison reflected a statistical perspective in his criticisms of Boring and did not find it plausible to base scientific conclusions on small sample sizes. Boring objected to Murchison's reliance on statistical summaries in order to draw inferences from his data. Boring considered such an act to be unscientific as it proceeded to make generalizations without an attempt to isolate causal factors.

Boring accepted small sizes in his experiments. Each subject's response was seen as a replication. It was important to find that under specified experimental conditions several people responded in a very similar way. Boring agreed that making generalizations was a difficult matter since there was not a specific outline to follow. He realized that there was a problem with small samples and accepted that it was difficult to know how many subjects would be required before the experimentalist could make a
reliable inference. He was convinced, however, that to increase sample size in lieu of increasing experimental control, could not lead to an understanding of the phenomenon under investigation.

What was convincing for the experimentalist was not convincing to someone who thought in terms of statistics and large sample sizes. At issue in this debate was which of two perspectives was scientifically acceptable in determining causal factors. The experimental approach was tied to the laboratory, the statistical to large samples. Each conceived of the generalizable finding in a different, even opposed, way. Where the experimentalist moved from the particular to the general case, the statistical psychologist studied large groups to arrive at principles and then applied these to individual cases. This latter approach was particularly well suited to a practical psychology where individuals were placed in a normative frame of reference.

The debate between Boring and Murchison reflects the conflict between two different perspectives on research procedure. These two perspectives also implied different requirements for institutional and financial support, both of which were limited resources. For instance, laboratory research required space and sophisticated equipment to operate effectively in the competitive scientific marketplace. Research based on statistical survey's required less space and less equipment. It did require a large number of researchers to carry out the studies as well as support staff to physically handle the data collected. In the remainder of this paper, I want to draw out the
politics of this debate in terms of its institutional impact and the bearing it had on who could best produce psychological knowledge.

The Politics

There is little question that E.G. Boring was a propagandist for experimental psychology. That he was active as a critic of the use made of statistical methodology in mental measurement is clearly documented in his publications on the subject which span the period between 1919 and 1960. The question arises, however, as to the reasons why Boring selected Murchison for criticism given the abundance of psychologists who employed statistics in much the same manner.

Boring stated in his 1926 note that Murchison provided a good example of the problems encountered when a researcher attempts to make a generalization. He also added that he hoped that his Note would provide, in a short form, an answer to Dr. Kelley’s earlier criticisms of his position on matters of statistical method. He wrote to Kelley after publishing the Note:

I am sending you a couple of reprints which you may or may not be glad to have. One of them, however, is a batch of notes the first one of which is an attempt to resolve our ancient controversy. . . . I apologize for lugging Murchison in so vigorously but it seemed too good a chance to accomplish an understanding with you.

He admitted in this letter that Murchison was "the type of
person" he had in mind when he first wrote his critiques of statistics. "[S]ome how I feel it in my bones," he added, "that you have no more use for Murchison than I."62

Some doubt can be cast on Boring's accusation that Murchison provided only a means to achieve an understanding with Kelley. Other reconciliatory avenues were open to him. Boring never published a direct response to Kelley's paper, despite the fact that he had written a rejoinder and Titchener encouraged its publication.63 Why did Boring, three years after Kelley's paper was published, publish this tract? An examination of the surrounding events, the history of the participants in these events, and the implications that may have occurred as a result of this note, belie the accusation that Murchison was incidental to the overriding purpose of Boring writing this 1926 note.

Institutional Interests

Both Boring and Murchison received appointments at Clark University, but at different times. At the invitation of G. Stanley Hall, Boring arrived at Clark in 1919 for a three year provisional contract. He was appointed as Professor of Experimental Psychology. If his work and productivity was regarded as satisfactory, he was to be appointed indefinitely.64 But he ran into some difficulties and departed in 1922 to accept an associate professorship at Harvard, a drop in rank but an increase in salary. Murchison gained his appointment at Clark shortly after Boring's departure.

Boring was not dismissed from Clark; he chose to leave. The events surrounding his choice tell us as much about
Boring as they do about the politics of being a part of a university.

Shortly after his arrival, Boring gained the impression that the new President of Clark, Wallace W. Atwood, was interested in sacrificing psychology to make room for an institute of geography. E.C. Sanford reassured him that it was unlikely that psychology would be dispensed with as it was a productive department. Yet, for Boring, all the evidence pointed to the contrary point. He complained to Sanford: "I have tried to get him [Atwood] up to the laboratory of 15 minutes to see us in our work and hear of what we are doing but have not yet succeeded."65

Boring wanted to feel more secure but even upon receiving an offer from Teachers College to join the faculty there, he turned them down. He confided to Sanford that he could not see himself connected with "that mantle", especially since he regarded "Teachers College as the antithesis of Clark and Clark as the ideal."66 But he could not trust "a mundane president", especially since the President had so much "autocratic power in determining the future of Clark." Therefore, before rejecting the Teachers College offer he spoke with Atwood in the hope of gaining some assurance that the research ideal would be maintained at Clark. He came away disappointed:

I came away from him marvelously discouraged . . . I had distinctly the feeling that he was anxious for me to go and that he felt here we were to start about getting rid of one of the unnecessary departments at Clark.67
As a direct result of this meeting, however, Atwood agreed to ask the Trustees to look into the matter of psychology's future.68

The Trustees delivered what seemed to be a positive judgement. They offered no changes to the present facilities, no reductions, and if reappointed, Boring would receive a substantial increase in salary. Sanford received the news with enthusiasm:

... right glad I am to know that all your troubles over the intentions of the new administration have been so happily put at rest. The letter from President Atwood is a regular charter for the department! We have nothing to do now but to go ahead and score!69

Boring for some reason was less optimistic. He stated that he would have been satisfied had the letter of support been "granted willingly" rather than "wring from him". He also sensed a "personal antagonism on the part of the president."70

Shortly after this letter of reassurance, President Atwood changed the rules governing the institute, "debasing" the research character of the university—according to Boring. What Atwood did was he amalgamated the university with Clark College, the undergraduate teaching college. Boring was incensed:

The only form of government we have this year is fiat accompli (416). .. My own stabilization is almost in the same class. I hate it, but do you wonder that I am inclined to fight the Devil with fire?71
Boring joined with other members of Clark's faculty to resist the administrative changes Atwood proposed. He thought that the changes that were being forced upon the faculty would take away from the research character of the university.

On March 14, 1922 the tension between the President and Boring came to a head over the issue of censorship. This was the day Scott Nearing spoke to the Students Liberal Club of Clark College. The message he delivered rang of socialism and drew heavily on the works of Thorstein Veblen. The issue discussed was the control of the university by vested interests. Atwood arrived late for the talk but before it was finished, demanded that the talk be stopped. According to one account, Atwood "ordered the janitor to extinguish the lights, and if that failed [to dispense the crowd], the police would be summoned." 

The students were outraged. Boring seized the opportunity to draw the attention of the nation to the autocratic rule championed by Atwood. He wrote to his parents:

Hell is popping at Clark University over free speech and presidential autocracy. . . Dr. Barnes and I have persuaded Croly of the New Republic to send Bruce Bliven; he is coming Thursday. . . I am working on Cattell.

Boring allocated space for Bliven in his laboratory, perhaps a politically insensitive move given the strained relationship between himself and the President.

In light of these circumstances, Boring was certain that
his reappointment was in jeopardy; Bliven told him as much. More now than ever, he searched for a new academic residence. Harvard came through with an offer, an offer Boring attributed to his work on the nature of scientific inference and statistics.

Due to circumstances outside of Boring's control, the press learned of his intention to take up the Harvard offer. This was unfortunate as he still wanted to bargain for his reappointment, if only to have the "right to resign." The following morning the paper printed the headline: BORING QUITS AS RESULT OF NEARING CASE. The afternoon edition printed the headline: BORING AND PRATT NOT WANTED SAYS ATWOOD.

Muckraking: Clark University, its Administration and Murchison

Boring was bitter about having to leave Clark. With its ties to the development of psychology in America, the fact that it was modeled on the Continental research tradition, Clark was where Boring most wanted to be. He perceived that Atwood's policies were destroying the research zeal at the university. Boring thought that experimental psychology could not survive under such circumstances, and while he was on faculty at Clark he fought the new administration. He wrote to a friend bemoaning his lack of productivity during his last year at Clark: "My chief energies in 1921-1922 went into fighting the new system at Clark University, which let me leave branded as a Red."

His difficulties with gaining funding for a suitable
laboratory at Harvard, the stress of establishing a research program in a department dominated by philosophers, was seen against the backdrop of what was available at Clark. One can sense his bitterness when he wrote to Sanford:

... the main cause of my fatigue is the tremendous difficulty in shaping Harvard into a place where I have the same opportunities for psychological development and productivity as I had under the old regime at Clark. This is what the trustees, and I think you too, will never realize, namely, the magnitude of the scientific opportunity which was destroyed.81

Boring was set to expose this destruction—show it up as the result of an administration that was insensitive to the conditions required for scientific research.

Boring feared that under the new administration if psychology was going to prosper it would have to turn away from experimental, laboratory research. What fed this fear was Boring’s belief that Atwood was bored with experimental psychology. When he had tried to advertise for graduate students in one of the psychology journals, Atwood refused to financially support the enterprise.82 The intentions of the new administration were proclaimed loudest, as far as Boring was concerned, when they hired Carl Murchison, a mental tester, as his replacement. Furthermore, Sanford was approaching retirement, and on retirement, Murchison would take over as department head.

Boring did not recommend that students should go to Clark for their Ph.D. He did not support Sanford in his
claims that the department had been "put on its feet again". Indeed when Sanford wrote to Boring requesting assistance in gaining for Clark the nomination that the APA meeting be held there, he refused. Boring was a member of the APA council and was in a position of influence. He wrote to Sanford that there were "two pretty obvious principles" which prevented him from supporting the nomination. The first was that since Sanford would be retiring, that would mean that the APA would gather when Murchison was taking control of the department. He wrote: "... the character of a department is determined by its head and that every head starts over. ... Murchison will inherit practically nothing from you." His second reason was that Murchison was an unknown. He explained his concern to Sanford:

... it is productivity that counts. Since Murchison is young and unknown, his own productivity will be watched at first, and if this brings him graduates or he gets them in other ways, then the productivity of his laboratory.83

The upshot of Boring's comments were that Clark would have to demonstrate that it had returned to its former stature. As far as Boring was concerned an institution had to be judged by the quality of the knowledge it produced. He continued his letter to Sanford saying:

... any other method of advertisement, like bringing the APA to Worcester, seems to me to be artificial and unlikely to produce results. It is, if I may say so strong a thing, almost Atwoodian.84

It was clear that Murchison was going to be watched. On
November 22, 1924 E.C. Sanford died and Murchison became head of the department. In the same year G. Stanley Hall died and Murchison became, as a result of his position as head of the psychology department, the editor-in-chief of Pedagogical Seminary. In light of such circumstances, it is highly unlikely that Murchison was drawn into Boring's 1926 note because he just "happened" across Murchison's research and saw in it an opportunity to make amends with T.L. Kelley.

One more circumstance testifies against Boring's assertion that Murchison was incidental to his Note. Murchison was not yet a member of the APA. Boring was on the APA council which recommended nominations for membership. Boring reviewed Murchison's research and, as a result, wrote his 1926 note, "Scientific Induction and Statistics." The note was a direct action to prevent Murchison from being nominated to the APA. Boring wrote to H.D. Kitson:

Somehow or other I have the impression that you are interested in knowing why the council of the APA deferred action on Murchison's nomination for membership until his publications could be examined more thoroughly. The action of the council was unanimous, and I was not alone in the discussion that looked toward this end. I have no right to speak for others, but the doubts that I raised before council are the doubts that I raised in this note. I think you will agree that publication of this kind is nothing short of scandalous.\textsuperscript{85}
Reviewing the publications of prospective nominees was common practice; writing and publishing critiques of such a review was not. Some took exception to Boring's procedures. One member of the council wrote to Boring:

I am wondering whether you are not a little severe in your conclusions that because of these defects in his [Murchison's] work he should therefore be kept out of the Association. I mean that I am rather afraid that if we went over the work of every individual there would be many who would be guilty of such lapses.86

**Professional Politics**

Murchison's application was exceptional and Boring gave it special attention. After all, Murchison was installed as a head of a psychology department at a major institution and was an editor of a major journal and was not yet a member of the APA. Boring wrote to June Downey, another member of council:

... we had a case, which will probably never occur again, of a young man just past his Ph.D., who has become editor in chief of one of our standard journals by sheer chance, ... and the strange act of President Atwood in appointing as head of that department a person who had not yet arrived psychologically.87

There can be little doubt that had Murchison been denied APA membership, Clark would have been in an embarrassing situation. By rejecting Murchison, a shadow would be cast
on the competence of Clark University's administration. That Atwood would appoint as head of a department an individual who was not acceptable to his professional peers would appear to be a hasty, ill-informed decision.

Boring used a methodological criticism to discipline Murchison and perhaps to discredit psychology at Clark. Other avenues of stricture were possible. In terms of strategy, Boring could have circulated among the members of the APA council a listing of the problems he found with Murchison's research and in this way achieve a moratorium on Murchison's membership. But he published the results of his review.

Boring not only publicly proclaimed the limitations in Murchison's handling of the data, he sent out fifty-three reprints to psychologists and philosophers who he thought were interested in the problem of induction. Seventeen of these reprints were sent to "statistical psychologists". Boring actively involved a community in his criticisms of statistics.

I think Boring's strategy was straightforward. Murchison bungled. To offer no account of variability in the assessment of data, and then to draw conclusions from these data based on proportions, was indefensible. Any statistician would accept that such an error is a serious oversight. So, Boring was confident that he would find common ground with the statisticians. The primary objective of his note—to argue that statistics are useful for description but not for prediction—was a point statisticians would object to. But how could they respond?
Could they write in defense of Murchison when he was clearly wrong? No one came forward to defend Murchison. Private protests were initiated, but no one published a defense of Murchison’s procedures nor a criticism of Boring’s argument.

Perhaps most psychologists judged that Boring and Kelley had already stated all that could be said on the point. Perhaps many, who were primarily concerned with the ‘doing’ of research, were confused as to what to think about statistics. Certainly this was the sentiment of G.M. Ruch when he wrote to Boring: "I have followed most of the discussion, and I am anxious to make up my mind as to the probable truth of the various conflicting assertions."

Conclusion: Policing a Profession?

If John O’Donnell is correct in his argument that Boring did all he could to prevent applied research from getting a foothold in American psychology, then it seems reasonable that Boring’s criticisms of statistical practices reflect this interest. In the case of Carl Murchison, Boring had the opportunity to prevent a particular form of applied psychology from gaining an institutional presence. I am suggesting that the debate between Murchison and Boring not be read as a personal squabble over methodology, though that is certainly one dimension. Rather, in blocking Murchison’s nomination to the APA, Boring was preventing a methodological perspective from gaining credibility as well as a substantial inroad to the structure of American psychology. If Murchison’s research was regarded by the APA as inadequate, so too would his direction of the Clark
department and his editorship of Pedagogical Seminary be so judged.

In the cases of Carl Brigham and Ben Wood, a similar pattern of interaction with E.G. Boring can be seen. In each case Boring used both his position in the APA, his influence with editors, to gain advantage in displaying the methodological short-comings of his mental-testing colleagues. In the case of Brigham and Murchison Boring used a criticism of their use of statistical methodology to point out their lack of scientific rigor. In the case of Ben Wood, well, it is difficult to assess just what was going on.

Certainly Boring was instrumental in preventing Wood from becoming a member of the APA and in this capacity he received criticism from both Terman and Kelley. We see in Boring’s dealings with Murchison a similar pattern of censure--he was bound to keep him out of the APA. He was unsuccessful in his attempt and many private protests were issued. In Murchison’s case there was more to be gained by discrediting him. Some of this was certainly motivated out of personal concerns, but Boring was also aware that this was one way of perhaps regaining Clark for the experimentalists. In this aim Boring was also unsuccessful.

Methodological debates need not be wholly concerned with matters of procedure. Even if it is granted that they are, there is reason to accept that matters of procedure include interests that serve the ambitions of particular professional communities. Indeed methodological debates correspond to different ideologies of quantification, and
these ideologies are the property of competing groups.
NOTES


5. Boring to Lashley, May 27, 1924.

6. Ibid.


8. Boring to Karl de Schweinitz, Jan. 4, 1924. Harvard Archives, Boring Correspondence.

10. See John O'Donnell, "The 'Crisis of Experimentalism' in the 1920s: E.G. Boring and his uses of Historiography." *American Psychologist* (1979) 34 289-295. His account is difficult to improve on and is convincing. The quote from Langford is taken from this paper, p. 293. The information in this paragraph is taken from O'Donnell's paper.

11. Ibid., O'Donnell mentioned that there was interest in having Titchener as APA President but Titchener would not accept the nomination unless it was unanimous. Boring's activities with respect to getting Bentley to run was revealed in a letter he wrote to Knight Dunlap, Dec. 14, 1922.


22. Ibid., p. 240.

23. Ibid., p. 240.


27. Terman to Yerkes, April 10, 1923. Stanford Archives, Lewis Terman Papers.


29. Boring to Terman, May 26, 1926. Harvard Archives, E.B. Boring Correspondence. On September 22, 1925 he wrote to Terman saying that he was working on this project. "Wouldn’t it be pretty," he remarked, "if eventually we should find simple fundamental speed differences, of the order of physiological simplicity of the rate of conduction of the nervous impulse, which has been shown recently to have surprising variability."


31. Ibid., p. 4.

32. Ibid., p. 254.


35. All of these quotations are taken from a letter Boring wrote to Kelley, February 27, 1923. Harvard Archives, Boring Correspondence.


38. Terman to Kelley, July 16, 1920. Stanford University Archives, Terman Correspondence.


40. This is an odd professorial title but Columbia College (a male college in the Columbia University system) insisted upon entrance examinations. In 1909 Columbia University appointed a professor whose chief duty was to function more or less as an admissions officer, judging the suitability of candidates for entrance into the college. I expect that Ben Wood served in this capacity while at Columbia College. For a brief history of Columbia College see Edwin Slosson, Great American Universities, New York: Macmillan Company, 1910, 442-473, in particular pp 459-461.

41. B.D. Wood, Measurement in Higher Education, Yonkers: World Book Company, 1923, see "Editor’s Note".

43. Boring to Kelley, Feb. 27, 1923.

44. Terman to Kelley, Jan. 28, 1923. Stanford Archives, Terman Correspondence.


47. Ibid., p. 303.


49. Ibid., p. 461. The quote was taken from Kelley’s criticism of Boring’s two papers, "Mathematical vs. Scientific Significance," (Psychological Bulletin 1919 16 335-338) and "The Logic of the Normal Law of Error in Mental Measurement" (American Journal of Psychology 1920 31 1-33). Kelley’s paper was, "The Principles and Technique of Mental Measurement," American Journal of Psychology (1923) 34 408-432. Kelley took exception to Murchison’s use of this quote without first seeking clarification and permission. Kelley wrote to him saying: "I regret very much that you have quoted with the context of my passage . . . In the context it was plain that these remarks referred to method and not to Boring. Stated so abruptly as you have done lends a personal implication which was not intended." Kelley to Murchison, July 10, 1926. Harvard University Archives, E.G. Boring Correspondence.


52. Ibid., p. 303.

53. Ibid., p. 303.


55. Ibid., p. 306.

56. Murchison, "An Answer to Dr. Boring," op. cit., p. 460. All the underlining was present in Murchison's paper.

57. Ibid., p. 460.

58. Ibid., p. 460.

59. E.G. Boring, "The Sensations of the Alimentary Canal," American Journal of Psychology (1915) 26 1-57. It should also be noted that this study was part of a series of studies on the alimentary tract. See Boring, "Processes referred to the Alimentary Tract: A Qualitative Analysis,"


61. Perhaps the most convincing demonstration of this has been provided by John O'Donnell, "The Crisis of Experimentalism in the 1920s: E.G. Boring and his Uses of History," American Psychologist (1979) 34 289-295.

62. Boring to T.L. Kelley, April 26, 1926. Harvard University Archives, Boring Correspondence.

63. Boring wrote to Kelley that "he was sufficiently piqued" by his "detailed criticism" that he "wrote a rejoinder." (Boring to Kelley, Oct. 6, 1923). Earlier in the same year Boring had written to Terman about his "row" with Kelley: "At present there seems to me to be only thirty-one things the matter with it; I shall probably find more later." (Boring to Lewis Terman, March 26, 1923.) For Titchener's support and encouragement to publish a response to Kelley's criticism see the letter Titchener send to Boring, March 22, 1923. Titchener provided advice on how Boring might proceed with the argument. Harvard University
Archives, Boring correspondence. I have been unable to locate the manuscript and it was never published.


65. Boring to Sanford, Nov. 10, 1920. Boring Correspondence, Harvard University Archives.

66. Boring to Sanford, Feb. 5, 1921.

67. Boring to Sanford, Feb. 15, 1921.

68. Ibid.

69. Sanford to Boring, March 25, 1921.

70. Boring to Sanford, Feb. 25, 1921.

71. Boring to Sanford, April 20, 1921. When Boring spoke of fighting the devil with fire he was referring to an incident concerning the advertising of the department. He tried to gain approval for advertising the graduate program in a psychology journal. Atwood blocked the proposal saying that the budget for advertising was expended. Boring wanted to go ahead by appropriating funds from the laboratory account but Sanford advised against such action suggesting that it would be seen as a "gross lack of consideration" and might undermine the confidence that the Trustees had shown to the department. See Sanford to Boring, April 18, 1921.

72. Boring referred to this in his autobiography and claims that they came to be called "the disaffected professors", who resisted changes that they regarded as debasing the
principles that the University stood for. See *Psychologist At Large*, op. cit., p. 35-36.


74. Ibid., p. 134.

75. Boring to "Family", March 22, 1922.

76. See *Psychologist at Large*, p. 36.

77. Ibid., p. 36.

78. Boring made such comments twice in his autobiography, pp 32 & 114.

79. See *Psychologist At Large*, p. 38.

80. Boring to Karl de Schweinitz, April 24, 1923. Boring Correspondence, Harvard University Archives.


82. Boring to Sanford, Jan. 16, 1920.

83. Boring to Sanford, Jan. 8, 1924.

84. Ibid.

85. Boring to Kitson, April 22, 1926.

86. Rudolph Pintner to Boring, May 3, 1926.

87. Boring to J. Downey, April 28, 1926.
88. Boring reported this to G.M. Ruch in a letter written July 15, 1926.

89. Ruch to Boring, July 12, 1926.
Conclusion

This thesis has been concerned with the identification of competing research traditions within psychology and their assimilation of statistical methodology. Specifically I examined the experimental and correlational research communities with an aim to display their different interpretations of statistics.

The experimentalists, represented by Edwin G. Boring, interpreted statistics strictly as a descriptive tool. To meet the conditions whereby statistics could serve in an inductive fashion was thought to be so restrictive that such circumstances seldom, if ever, arose. The correlationists, represented by T.L. Kelley, saw statistics as an inductive technology. For them, statistics provided a means through which causal networks could be identified and new data brought to light.

It would be wrong, however, to credit the competition among opposed ideas as being responsible for the conflicts between these research traditions. Rather, the conflict stemmed from differences over how science should be practiced, how labor within the research community should be organized, how knowledge should be generated, who should generate it, and who should apply it. Science activity, in other words, is constitutively social.

In this thesis I argued that if science is constitutively social, then all science is applied science in the broadest sense of that term. That is, science is a process and a product of social organization. The knowledge
generated by a research community inherently reflects social agreements that serve the interests of scientific groups.

Further, the boundary-talk between pure and applied science is not be to be accepted at face value but needs to be understood and explained.¹ That the boundary setting apart pure from applied science was often discussed by psychologists in the past should suggest questions, not provide answers - such as who were applied and who were experimental/pure psychologists. The question that arises addresses the issue of what was gained by a particular research group in talking about the pure/applied distinction in the way they did. In this work I interpret methodology as reflecting socially agreed upon boundaries on research activities. These boundaries in turn, reflect social, cultural, intellectual and technological interests as they pertain to the role of psychological science in society.

In this thesis I have treated methodological debates as involving broader issues concerning not only disagreements over knowledge generating activity, but also over who should be allowed to generate and apply psychological knowledge. Thus I organized the chapters to draw together three dimensions: (1) the social demands made on psychologists during the 1910s & 1920s and the nature of their response; (2) the technical debates over the use and interpretation of statistical methodology; (3) the strategies and tactics of week-to-week politicking within psychology and how this interacted with methodological debates to shaped who could gain access to the profession.

In the first two chapters I drew upon the social turmoil
of the times to suggest that Americans perceived their society as being corrupt and requiring "expert" guidance in reestablishing social stability. But the definition of expert was loose. Clearly the science community had the most to lose by not establishing who should be considered an expert.

Two models of expert were presented, each representing different notions of applied science. The first I called the "objective arbitrators". These experts maintained the central role of the scientist in the production of scientific knowledge. That is, the scientist was the arbitrator of what constituted a generalizable fact. The second expert I called the "manager of methods". These researchers interpreted their methods as an inductive technology. Statistics were quickly assimilated by this group as a methodology that separated inference from opinion. The demand for statistical training by psychologists increased dramatically, as did the textbooks.

In the next three chapters I examined conflicts over the use and interpretation of statistics. By connecting experimental psychology with the "objective arbitrator" model and correlational psychology with the "manager of methods" model, I attempted to illustrate that debates over methodology were concerned with broader social issues. I attempted to show that interpreting statistics as being descriptive - and the psychologist as an "objective arbitrator" of the facts - traditional, university based, laboratory psychology could maintain its footing in post World War I institutional growth. If however, this approach
to research lost its appeal as providing accessible knowledge for practical ends, it risked losing credibility. The "objective arbitrator", the 'life-style' of research associated with this model, would be lost.

Interpreting statistics as providing an inductive technology led psychologists to believe that they could abandon the laboratory and conduct research in the 'field'. They interpreted statistics as being practically predictive instruments which allowed them to statistically control and interpret the meaning of variability. Also, since it was the reliability of the methods that were emphasized, and not that of the researcher, the ways of doing research was reorganized.

This reorganization reflected the corporate model of business management. Graduate students were taught how to use statistical methods, they were sent into the field to collect data, and the head researcher wrote the report. Also because these "managers of methods" worked in the 'field' among other non-psychologist professionals, they promoted statistics as a way to secure agreements. If the critiques of statistics were effective, they could not be used in a conciliatory fashion. The statistical critiques of the experimentalists (and other objectors), threatened the cogency of their entire research program. Thus, their response to these critics was terse and unsympathetic. They generally portrayed their opponents as being either illiterate or misinformed about statistics.

Because the credibility of expertise of each research tradition was being undermined by the other, the debates
over statistical methodology were heated and defensive. It was after all, life-styles of research practice that were being challenged. This brings me to an important difference that emerges from Chapters 3 & 4.

It is clear that the experimentalists and the correlational psychologists drew battle lines. Ruml was an industrial/applied psychologist. He was also critical of how mental testers used statistics. Ruml appeared to adopt an "objective arbitrator" model of expertise. Nevertheless, his "life-style" research context was quite different from that of the experimentalists and the correlationalists.

The applied psychologists who set up as consultants to business and industry were in a situation that was at odds with the two university based research groups. To survive financially, they had to sell their product. In a sense this was also true of both the experimentalists and the correlationalists, but less so, because of the financial cushion provided by the university.

Beardsley Ruml, Robert C. Clothier, and Walter Dill Scott all served industry directly as consultants. Together they formed the Scott Company, and offered to provide industrial firms personnel consulting. It was during these years that Ruml penned his criticisms of statistics. They developed several specific tests, and, as Hilgard (1987) noted, they were recognized for their rejection of the concept of the "square-peg-in-a-square-hole" approach to personnel management. Indeed Ruml was noted for his recognition of the "man-machine relationship and for the concept of the 'worker in his work'." In this research
they endeavoured to study the applied environment in detail and followed more closely the experimental approach in terms of their efforts to identify and control that which caused variability. They worked more from an "objective arbitrator" model. The Scott Company failed after three years and folded in 1922.

In 1917 the Carnegie Institute of Technology established a unique program in applied psychology. Walter Van Dyck Bingham established the program and set it up as a consulting firm, more or less. This program was disbanded by 1924. It also assumed a more "experimental" research perspective.

Both the Scott Company and Bingham’s efforts at Carnegie Tech, illustrate an important deviation from the type of applied research carried on by the mental testers. They had to sell a service that brought a dollar return to industry; the mental testers and the experimentalists were not so constrained by fiscal objectives. What it amounted to, I think, was that individuals like Ruml, Bingham and Scott sold a product to industry - this could be tested for its economic contribution by a company. For the most part, they failed to provide industry with a reasonable dollar return and their consulting operations folded.

The experimentalists and correlationalists (mental testers, for the most part) sold their knowledge generating process, not its product. Thus mental testers competed with the experimentalists in the university structure, and not so much with the industrial consultants in the business environment. Therefore we find the correlational
psychologists seeking to establish - and chapters three through five show this - that their methods were scientific and, on close inspection, followed the very cannons of scientific practice that motivated the experimentalists.

The mental testers, then, sought parity with the experimentalists. They sought to establish their operative base within the confines of the university system, though they used the university in ways quite different from the experimentalists. Correlationalists sought to establish their presence throughout the professional domain of psychology.

Since mental testers and experimentalists alike were selling more the processes through which scientific knowledge should be generated, rather than the products of this activity, their methodological differences were drawn in sharp focus. It follows that their methodological debates were politically intense.

The final chapter established this point. It focused on the push-and-shove politicking that went on in psychology. Of course the reason this has a bearing on the present study was because politics and methodology interacted. The assimilation of statistics into psychology therefore represents changes to a social ordering of research practice. We witness social revolution in methodological debates, not ideational changes pulled along by the rational criteria of theory appraisal.
Epilogue

Although I have written the conclusion to this thesis, I do not feel the topic has approached closure. There are so many materials, so many debates over statistics, that I have left out of these six chapters.

I have not carried these statistical arguments into the 1930s. Here we find both a growing disenchantment with mental testing and the development of factor analysis. We find disputes between T.L. Kelley and Charles Spearman over the calculation of probable errors. Also we see S.S. Stevens picking up Boring's mantle and challenging psychologists to examine measurement theory. A conflict between Stevens and L.L. Thurstone forms at this time and its pattern fits nicely with some of the background issues I developed in this study. A hint of their conflict can be found in a letter from Stevens to Harold Gulliksen where he writes:

To me, Thurstone proposes to measure by processing variability. That is, the main procedure, the basic operation. Take variability away and there is no measurement under this scheme. The contrasting approach proposes to measure by ignoring variability completely - except to try to reduce it by good experimental techniques. This basic and elemental difference matters profoundly, it seems to me.3

I think it is appropriate to discuss the debates over factor analysis in a thesis that purports to be a study of the assimilation of statistics into psychology. These debates grew more intense in the 1930s and the conflict...
between Kelley and Thurstone is quite visible. Of course the general objections to factor analysis by Edwin B. Wilson in the late thirties is important, as is Kelley's response to him.

Yet to include all of this material would have required further evaluation of American society as the depression drew near and a more elaborate depiction of the politics of the APA in these years. The APA faced in the 1930s the problems of a growing membership and a lack of employment opportunities for psychologists. We find the growth of professional groups outside of the APA, such as "Social Psychologists for the Study of Social Issues" and the "Psychonomic Society", both of which expressed their opposition to what was going on in the APA. These issues clearly had a bearing on methodological debates. But to deal with these materials, the integration required, would have extended this study well beyond one volume. This thesis then, provides a first step toward a fuller study of the assimilation of statistics into psychology.
Notes

1. I am not suggesting anything new here. The relationship between pure and applied science, between science and technology, is a topic that is receiving increasing attention from sociologists, historians and philosophers of science. That the distinction is blurred is generally accepted. See, for example, Edwin Layton, "Mirror-Image Twins: The Communities of Science and Technology in 19th-Century America," Technology & Culture (1971) 12 562-580. More recently Barry Barnes and David Edge provide a succinct discussion of the problems of the pure/applied science distinction. They set out to portray it as a division of labor, with the associated problems that emerge when we (as sociologists, historians and philosophers) have not provided in the literature a sufficient understanding of the basis of credibility. See Science in Context (Milton Keynes: Open University Press, 1982), pp. 233-249.


BIBLIOGRAPHY


Babbitt, Irving. (1908) Literature and the American College Boston: Beacon Hill.


Birnbaum, Lucille C. (1955) "Behaviorism in the 1920s,"
American Quarterly 7 15-30.
__________ (1920) "Apriori Use of the Gaussian Law," Science 52 129-130.
__________ (1923) "Facts and Fancies of Immigration," The New Republic 34 245-246.
__________ (1926) "Scientific Induction and Statistics,"


(1941) "Statistical Frequencies as Dynamic Equilibrium," Psychological Review 48 279-301.

(1950) "Learning vs. Training for Graduate Students", American Psychologist 5 162-163.


Resurgence of an Old Misconception," Psychological Bulletin 87 564-567.


of Chicago Press.


Holzinger, K.J. (1923) "An Analysis of the Errors in Mental Measurement," *Journal of Educational Psychology* 14 278-288

________ (1928) *Statistics in Psychology and Education* Boston: Ginn and Company.


(1918) "Correlation," Psychological Bulletin 15 114-122.


(1985) "Pragmatic Realism and the Macrosociology of Experiment", a paper presented to the British Society for the History of Science/British Sociology Association Conference on The Uses of Experiment, Bath, August 30-September 2.


(1921) "Empiricism versus Formalism in Work with Mental Tests," (Journal of Philosophy 18 393-398).


(1920) "The Need for an Examination of Certain Hypotheses in mental Tests," Journal of Philosophy 17 57-60.


(1921) "Intelligence and Its Measurement: A Symposium," Journal of Educational Psychology 12 143-144.


338

(1913) "Editorial: The Educational Efficiency Engineer." Educational Psychology 4, p. 244.


London: George G. Harrap & Company LTD. (British edition of the 1916 American printing)


(1922) "The Great Conspiracy or the Impulse Imperious of Intelligence Testers, Psychoanalyzed and Exposed by Mr. Lippmann," *The New Republic* 32 (December 27) 116-120.


(1913) "Educational Diagnosis", *Science* 37 133-142, p. 142.

(1921) "Measurement in Education", *Teacher's College Record* 22 371-379, p. 371.

(1910) "The Contributions of Psychology to Education", *The Journal of Educational Psychology* 1 5-12, p. 12.


Townsend, J.T. & Ashby, F.G. (1984) "Measurement scales and
statistics: The misconception misconceived."
Psychological Bulletin 96 394-401.

Thurstone, L.L. (1917) "A Method of Calculating the Pearson Correlation Coefficient Without the Use of Deviations."


