Advanced

Quality Control Theory

for

Training and Education

A Guide to Optimizing Training and Education Efforts

Brad Heppler
ph. (360-292-2046)

In loving memory of my brother Kurt Heppler

Special Thanks for Editing:

Robert McConkie	B.S. English
Mack Finlayson	B.S. Bio-Chemistry
Wendy Finlayson	B.S. Geology MBA

Special Thanks for Equation Editing:

Hajime Natsume	M.S. Mathematics

Special thanks to Texas A&M for allowing me the freedom to invent.

About the Author:

B.S. Natural Resources	Utah State U.
B.S. Management	Weber State U.
M.Ed.	Texas A&M

I worked in the Silicon Semiconductor Industry for the last 15 years as a Lead, Process Technician and Trainer. This industry simultaneously faces increasing complexity, ever tightening specifications and decreasing prices. During this time I became interested in process controls and especially statistical process control. While at graduate school, I became interested in how to develop statistical process control techniques that are useful for the Training and Education industries. This book contains my thoughts on this subject.

Table of Contents

Chapter 1: Introductory Concepts

Chapter 2: Observations as Measurements

Chapter 3: Tests and Examinations as Measurements

Table of Contents (cont.)

Table of Contents (cont.)

1

Introductory Concepts

1.1 Quality Control Defined

This is a book about quality and how to control quality through deliberate actions on the part of the professionals developing and implementing the instances of instruction available at an organization. Quality in this context is not the highest level of excellence that can be achieved regardless of cost. For our purposes, quality is the level of excellence that first, meets the expectations of the organization and the marketplace and second, meets these expectations at the lowest cost. In this definition of quality, the organization may also be the marketplace, as is the case for in-house instruction in safety or other specialized knowledge. Cost efficiency is an integral aspect of quality and is therefore a key focal point in the process of improving the efficacy of the instruction available. Quality control methodology involves the repeated application of a series of statistical techniques that have proven useful historically for quality improvement activities. When quality control methodology is applied to the courses offered by a specific organization, the result is a quality control program. The quality control program becomes the summation of the organization's efforts at collecting and using measurements in order to establish, maintain and improve the efficacy of instruction. Quality control theory favors no particular learning philosophy and is only directed towards aspects of how, what, where and when measurements are collected on instruction and upon how these measurements are leveraged in order to maintain and improve instruction.

1.2 Economics of Quality

The basic goal of a quality control program is to strike a balance between the costs of quality and the value of quality at all levels of the organization. Of course upper-level management involvement is crucial for success but because information concerning the costs and value of quality are spread throughout the organization, a restated goal is to determine the optimum balance between the costs of quality and value of quality at the value added level of the organization. Many quality costs are avoidable at little expense, for example, by relaxing an overly tough criteria or clearing up misinterpretations of expectations. Other costs associated with quality are avoidable through considerable, yet still economically viable, investments in changes to an organizations instructional offerings, expertise and technology. An important point is that it is only possible to eliminate all of the costs attributable to quality at great expense.

In general, the higher the level of the agreement between the organizations goals and the instruction in place to help achieve these goals, the lower the total cost of operating the organizations instructional offerings. Precise duplication of a course on separate occasions, at possibly separate locations using different instructors also helps minimize instructional costs because precise duplication is the best way to eliminate the rather large and uncontrolled number of variables associated with anecdotal modifications to a course from one instance to the next. Anecdotal modifications typically result in additional costs associated with reassessment and remedial instruction and add little enduring information to the institutional memory.

Typically, the optimal level of agreement between the organizations goals and the instruction in place to help achieve these goals is achieved when further improvements will offer economically insignificant gains. After the level of agreement is optimal, the goal of the quality control program becomes maintaining the status quo through the monitoring function. Establishing and maintaining quality is accomplished through extensive monitoring of the ongoing quality of the instruction provided by an organization. When measurements are graphed as one or more time series, the monitoring function generates signals that indicate when quality has slipped and requires action in order to restore quality to the previous level. A second role of the monitoring function is to validate improvement efforts made to instruction using signals such as a persistent upward shift in one or more relevant time series.

1.3 Quality Control Program

For the purposes of a quality control program, a course is comprised of two components, the instructional process and the content to which the instructional process is applied. The instructional process consists of the complete collection of resources necessary to bring a course into being, including the delivery of each instance of instruction. The second component of a course is the content which consists of the information presented during a course. During the design of both the instructional process and the content, a control plan is developed consisting of a production control plan and a process control plan. The production control plan contains all of the information and references to outside information necessary to completely define how each instance of instruction comprising a course is to be produced. Production, in this sense, is defined as the complete sequence of events necessary to bring each instance of instruction into being, including details concerning the actual delivery of the content. Corresponding line for line with the production control plan is the process control plan which contains information regarding how the execution of each aspect of the production control plan is to be verified. The ability to reliably reproduce an instance of instruction each time the instruction is offered is critical for obtaining a high level of quality.

The production control plan indicates which strategies, also known as process controls, must be in place in order to insure the reliable reproduction of a course. These strategies are also deployed to varying degrees during efforts to improve a course. There are three basic types of process controls: Instructor dependent controls, Administrative controls and Engineered controls. Instructor dependent controls are the weakest form of controls and basically involve instructors adhering to the production control plan with minimal management oversight. Administrative controls consist of reconfiguring the existing mix of resources in ways that are perceived to be more effective. Instructor dependent controls and Administrative controls, whether used individually or in concert are only partially effective. These two control strategies also generally require additional man hours for implementation and verification which increases the costs associated with a course. Engineered controls are the strongest form of controls and consist of implementing the production control plan in ways that preclude modification by facilitators, instructors or managers. Web based training is an example of using engineered controls to insure that students are exposed to content in a uniform way. When rigorous adherence to the production control plan is necessary, and this should always be the case, engineered controls are always viewed as superior to instructor dependent and/or administrative controls.

Implementation of a quality control program is a complex and evolving undertaking which requires the consistent, persistent and recurrent application of the interrelated steps prescribed by quality control methodology. The most efficient, and thereby the lowest cost, point in time to apply quality control methodology to an instance of instruction is before widespread implementation. For existing courses, inclusion in the quality control program is usually based upon organizational priorities or by using some method of assigning value to a particular course. Once a new or existing course is selected for inclusion in the quality control program, the format of the production control plan and associated process control plan may differ from one organization to another. Our focus in this text mainly concerns the portion of the process control plan which defines how the efficacy of instruction is to be verified.

The first step in this verification process is to determine at what points in time to collect measurements. Once the measurement points are identified, the instance of instruction occurring between two consecutive measurements is called a Sequence of Instruction, abbreviated SOI, and the SOI becomes the basic unit of instruction for purposes of the quality control program. An exception to the definition of an SOI occurs for a course comprised of a single SOI or for the SOI that is the first in a collection of SOIs because typically no relevant beginning measurement is available.

The simplest case is when a course consists of a single SOI as a result of having only a single measurement point. Often a course contains several SOIs or contains SOIs that depend upon the knowledge contained in one or more earlier SOIs for success in understanding the current SOI. When SOI's are interrelated, the ability of the collection of SOIs to produce adequate performance must be verified for each individual SOI as well as for the collection of SOIs as a whole.

1.4 Measurements and Scales

Measurements are the result of the operation of a measurement system. There are three basic classes of measurement systems used to measure the products of instruction. The first consists of measurement systems that rely upon responses provided by students to instructor generated questions in order to demonstrate their level of knowledge concerning the content of an SOI and these will be referred to as instructor generated examinations or examinations. The second class of measurement systems also relies on student responses to questions designed to assess their level of knowledge but this type of measurement system depends on questions developed and selected using psychometric methods. When the questions are developed and selected using psychometric methods, these will be referred to as standardized tests or tests. Standardized tests are typically developed to assess knowledge of a broadly defined body of knowledge known as a domain. A domain is so large that only a fraction of the possible questions can be included on a single test. Psychometric methods allow the best set of questions to be selected from a larger pool of questions in terms of delineating between students having different levels of knowledge of the domain. The third class of measurement systems rely upon observations. Among these are observations collected by instructors concerning a student's performance and observations collected from students concerning the quality of instruction. Another type of observations occurs when trained observers seek to determine aspects of the instructional process that may need further development.

At a more fundamental level, individual measurements are classified into different types based upon the scale from which the measurements arise. Measurements arising from a nominal scale impose categories on a set of measurements but do not contain information regarding the rank order of the measurements. An example of a nominal scale occurs when a class is divided into noncompetitive teams. Measurements arising from an ordinal scale impose categories upon a set of measurements that also contain information regarding the rank order of the measurements. However, an ordinal scale contains no information concerning the distance between measurements. For example, teams ranked either first, second or third in a competition is an example of an ordinal scale. Measurements arising from an interval scale impose categories and provide information regarding order but in addition, the distance between measurements has meaning. For example, if Team 2 scored 96 points and Teams 1 and 3 scored 66 and 61 points, respectively, clearly Team 2 exhibited superior performance, where as, the difference between the performance of Teams 1 and 3 is trivial. In a true interval scale, the interval between any two rank positions is infinitely divisible. The property of infinite divisibility between ranks positions are seldom met by instructor generated examinations but long standardized tests may approach an interval scale in detail. Measurements arising from a ratio scale have all of the properties of an interval scale but also have an absolute zero point representing the complete absence of the characteristic being measured. The Kelvin temperature scale is a ratio scale because 0 degrees Kelvin indicates the complete absence of heat while the Fahrenheit temperature scale is an interval scale because 0 degrees Fahrenheit does not indicate the complete absence of heat. Interval or ratio measurement scales are often referred to as continuous measurement scales.

The type of measurement scale from which a measurement arises has implications concerning the amount of information contained in each measurement. Measurements based upon either an interval or a ratio scale contain more extractable information than those based upon an ordinal scale precisely because the differences between rank positions has meaning. For example, a forty-question examination, on which everyone misses ten questions or less, produces measurements based upon an ordinal scale having eleven categories rather than on an interval or ratio scale. When measuring the products of instruction, examination scores that reduce to an ordinal scale may be the only alternative. However, many times it is possible to create a measurement system that is truly interval or ratio or at least approaches an interval or ratio scale in detail. When using Grammatik or other text analyzers to score an essay, the resulting indices may approach an interval scale in detail. For example, if an essay contains 17% finite verb phrases, scores 16 of 100 for sentence complexity and 34 of 100 for word complexity, in addition to a letter grade or point score assigned by the instructor, the combination of these measurements provides much more information than a letter grade or point score alone. An additional advantage of using software, such as a text analyzer, is that the same measurement is produced regardless of whom runs the analysis program. This ability to reproduce measurement from one instance to the next or between different persons performing the measurements are defining characteristic of high quality measurements having potentially large amounts of extractable information.

A great challenge in the field of educational and training assessment today is to devise measurement systems that produce measurements arising from either an interval or a ratio scale that have high levels of extractable information.

1.5 Economics of Measurements

Much attention must be paid to methods for insuring the quality of all measurements. When assessing the quality of instruction, the current sophistication of measurement systems varies widely and this fact has made establishing historical levels of performance difficult if not impossible. The most common measurements rely upon students suppling answers to questions contained in instructor generated examinations that are also scored by the instructor who developed the examination. In this most common form of measurements, the summary of the results are also calculated by and disseminated by the instructor who developed the examination. A variation upon this system is to present an instructor-generated examination through a computer which automatically scores the examination as directed by the instructor. These types of measurements are uncontrolled and are typically of low or unknown quality where the results likely reduce to either a letter grade or a pass/fail determination of speculative value. Measurements having low quality create a situation in which an unknown number of students advance thru an SOI despite having an inadequate understanding of the content of the SOI. All of these factors when combined, point to measurements having low value per measurement for which the cost of each measurement is high. For these reasons, a significant number of the chapters in this text are devoted to the theory of measurements and methods of increasing the quality and thereby the value of all measurements collected by an organization.

Briefly, the value per measurement can be increased by migrating from uncontrolled measurements based upon systems such as grades or pass/ fail determinations, to carefully controlled measurements arising from a continuous measurement scale. A large part of this increase in value occurs as a result of controls on the development of measurement systems. Also, value increases because the interval between any two measurements has meaning.

The cost per measurement can be reduced by instituting passive data collection systems and by automating data analysis. In a passive data collection system, measurements are collected and recorded automatically at the time of measurement. In the absence of a passive data collection system, analysis of large quantities of measurements is laborious, in the extreme, and is rarely undertaken except under the most pressing circumstances. Passive data collection systems allow sophisticated analysis to be automated where the results of the analysis are made available to all interested personnel. The cost per measurement can be further reduced by collecting multiple measurements at each measurement opportunity.

1.6 Overview of Implementation

Figure 1.1 contains the five groups of statistical procedures necessary to implement quality control upon an SOI. In Step 1, a course is selected for inclusion into the quality control program and the SOIs that comprise the course are identified. In more complex courses, there are typically several SOIs. Several methods exist for selecting an SOI, a course or collection of courses for inclusion in a quality control program. One popular approach for selecting SOIs for inclusion involves ascribing a value to each SOI based upon the perceived importance of the SOI to the attainment of organizational goals. Techniques such as Forced Field Analysis or other forms of multi-dimensional decision analysis are often used to produce such a ranking. When success in one or more SOI's is dependent upon knowledge acquired during earlier SOIs, diagraming the network of all antecedent SOIs in relation to the current SOI is necessary. The SOI selected for improvement first is the one offering the highest potential gains for the network. For example, if poor performance in a particular SOI in Chemistry I causes difficulty for success in Chemistry II, the SOI from Chemistry I will require improvement before improvement of Chemistry II is possible. There are certainly many other methods for selecting an SOI for improvement but the main criteria is what makes sense for the organization. After an SOI is identified for inclusion in the quality control program, the instructional process for the SOI is defined.

In Step 2, three distinct classifications of measurement systems are discussed but others are possible. The goal here is to determine what measurement system will be used or is already in place for assessing each SOI selected for inclusion in the quality control program. The first type of measurement system is based upon observations. This type of measurement system is typically organized around an observational instrument, such as a questionnaire or upon counting or timing instances of the behaviors of interest. The accuracy of this type of measurement system is demonstrated by repeatedly measuring recorded or otherwise preserved samples similar to that likely to be encountered upon implementation. The objectivity of an observational measurement system is comprised of two components: Repeatability and Reproducibility. Repeatability is the ability of a measurement system to produce the same results for the same situation on the same occasion. Reproducibility is the ability of a measurement system to produce the same results, for the same situation, on separate occasions. Determining the reproducibility component of objectivity requires the use of recorded or otherwise preserved samples. If appropriate samples are not available, only the repeatability component of objectivity can be determined.

Observational instruments are often based upon Likert scales where the measurement process involves selecting numbers from a continuum to indicate the level to which a particular behavior or characteristic is present or absent. Observational instruments are developed using Psychometric Theory which is a collection of statistical techniques that seek to select the individual scales that maximize the objectivity of the observational instrument as a whole. Regardless of the origin or type of observational instrument a measurement system is based upon, an accuracy and objectivity determination is required before implementation. The second type of measurement system is based upon an instructor generated examination. The accuracy of the measurements resulting from this type of measurement system relies upon a demonstration that the sum of the major sources of inaccuracy are 3% or less. Determining the objectivity of an instructor generated examination requires that two equivalent forms of the examination be developed simultaneously. The correlation between the two forms is used to indicate the level of objectivity. The sum of all sources of variation attributable to the lack of objectivity must be 10% or less of the range of measurements included in the objectivity study. Simulation studies of decisions based upon statistical inference have shown that if the percent inaccuracy is greater than 3% or if the lack of objectivity is greater than 10%, incorrect decisions can result. A full treatment of the topics of accuracy and objectivity are included in Chapters 2 and 3.

Figure 1.1: Steps Required for Implementation

1. Select a Course
  A. Define SOIs
  B. Define Instructional Process for Each SOI

2. Define Measurement System for Each SOI (*Select Either A, B or C*)
  A. Based Upon Observations
    1. Evaluate Accuracy
    2. Conduct a formal Objectivity Study
  B. Based Upon Instructor Generated Examination
    1. Evaluate for Accuracy
    2. Evaluate Objectivity
    3. Determine Minimum Level of Performance
      a. Identifiable Alternative Distribution, use Minimum Correct Procedures
      b. No Identifiable Alternative Distribution, use Distributional Specification Procedures
  C. Based Upon a Standardized Test
    2. Evaluate Psychometric Reliability

3. Evaluate Reliability (*Select Either 1 or 2*)
    1. For a collection of Independent SOIs use simple Reliability Model
    2. For a collection Dependent SOIs use Markov Chain Analysis

4. Conduct Short-term Capability Analysis (*Select Either 1 or 2*)
    1. Quasi-experimental Design
    2. Design of Experiments

5. Implement Monitoring
    1. Select Chart Types
    2. Develop Infrastructure for Passive Data Collection
    3. Implement Passive Data Collection and Analysis

Use of a measurement system based upon an instructor-generated examination requires a determination of the minimum level of performance. If an alternative distribution exists, the minimum number of questions that a student must answer correctly, on either form of the examination, is based upon the maximum acceptable probability of giving a student credit for completing an SOI when the student actually possesses inadequate understanding of the content. This is called the Type II error rate. The Type II error rate is a component of the percent inaccuracy. If no alternative distribution is identifiable, use the distributional specification procedures. When using a distributional specification, the minimum level of performance is not a single number of questions answered correctly but instead is the conformance of the score distribution for the SOI to the probabilities of a binomial distribution having an expected probability of a correct answer very close to 1.0. Calculating the number of questions that must be answered correctly and the development of distributional specifications for the score distribution are the topics of Chapter 4.

The third type of measurement system is based upon a standardized test. In this type of measurement system, the psychometric characteristics of the standardized test are of interest. The most common type of psychometric reliability, known as KR20 reliability, can be determined by administering the test to a large and representative group of students on a single occasion. This measure of psychometric reliability is then used to imply the level of objectivity of the measurement system without an actual demonstration of objectivity. Psychometric reliability of standardized tests is discussed in detail in Chapter 3.

In Step 3 in Figure 1.1, the reliability is determined for each SOI included in the quality control program. Reliability is not to be confused with psychometric reliability which is a type of objectivity. Reliability of an SOI or a collection of SOIs equals the probability that a student completing the SOI or collection of SOIs will perform at or above the minimum level of performance. There are two types of collections of SOI's important from a reliability standpoint. The first is a collection of independent SOI's that are broadly related but where success in each SOI is independent of successful completion of prior SOI's in the collection. This type of collection is known as a category of SOI's and a common example of a category of SOIs is a collection of safety-related courses each comprised of a single SOI. Categories of SOIs are defined by one or more persons in an organization based upon some type of decision rationale. The second type of collection of SOIs is a collection of dependent SOIs and this collection is known as a series of SOIs. Successful completion of a series of SOIs is dependent upon acquiring adequate knowledge of prior SOIs in the series. A series of SOIs is defined by analyzing the content of SOIs that are believed to contain antecedent knowledge. The reliability of a series of SOIs is evaluated using a simple matrix-based procedure known as Markov Chain analysis.

In Step 4, a short term capability study is conducted, in the environment and with the resources that will be available upon full implementation, for each new or improved SOI. There are two general types of study designs advocated here. The first is a quasi-experimental design in which a group of non-randomly selected students having mixed ability are exposed to the instruction and then measured. Success is achieved if all students score at or above the minimum acceptable score, if an alternative distribution is identifiable, or if the score distribution exceeds conformance requirements in the case of a distributional specification. The second type of short term capability study involves using Design of Experiments (DOE) methodology. In this type of study, groups of students are randomly selected from groups based upon categorical variables known by the organization to be predictive of performance. A variant on this theme is to form one or more groups from the students who have historically scored below the median. The status quo is used as a control group. The assumption here is that instruction that is effective for lower performing students will perform satisfactorily for most other students. This type of study is illustrated in Chapter 6. Other types of study designs, such as nested or factorial designs, are also useful.

A second use of short-term capability studies is for acceptance testing of instruction purchased from or provided by other organizations. It is imperative that outside developers of instructional product face the same high expectations of their products as for those developed in-house.

Step 5 activities all revolve around implementation of the monitoring function for certain, if not all, SOIs. This is the most visible portion of a quality control program but, as is obvious from our discussion so far, on going monitoring of an organization's SOIs is at the end of the implementation process for an SOI rather than at the beginning. The monitoring function involves the selection and development of charts, known as control charts, of various types. Implementation of passive data collection, analysis and monitoring is advocated at this step.

1.7 Concluding Remarks

An almost universal experience for organizations applying quality control methodology to an existing SOI, which was not initially developed under the rigor of a quality control program, is that the assumed reliability of the SOI has been dramatically overestimated. Upon encountering low reliability, there is a tendency to recoil from quality control implementation as if measuring the outcomes of existing instruction caused the results. Do not become discouraged, the road to excellence is long and challenging but the insights gained can be leveraged to the organizations advantage.

Cost efficiency is an integral aspect of a quality control program. If the reader is envisioning a quality control program that increases costs without parallel and multiple increases in efficiency the reader is in error. The idea is to select improvement efforts that offer the largest possible gains. Gains in this context means that all improvements implemented should be both statistically significant and economically significant. Statistical significance indicates that an improvement effort produced gains in excess of that likely to occur from random fluctuations. Statistical significance does not imply that the gains are economically significant.

As shown in Figure 1.1, implementation of a quality control programs requires that a series of statistical procedures be conducted on every SOI included in the quality control program. High quality measurements, ones having total inaccuracy of 3% or less and systematic error related to repeatability and reproducibility of less than 10% of the range of measurements expected in the implementation environment are of critical importance. Implementing a quality control program that lacks rigor will result in additional costs without realizing the expected gains. Lack of rigor is the overwhelming reason for failure of quality control programs.

2

Observations as Measurements

2.1 Introduction

Direct observations made concerning instructional practices or of student progress are probably the oldest forms of instructional assessment and continue to be in widespread use. Observational measurements can be collected on actual behaviors or upon the products of behaviors. Products of behaviors are physical items documenting the results of instruction such as writing samples. Measurements made by people, known as active measurements, are relatively expensive on a per measurement basis. Observations collected and evaluated by a device or series of devices, known as passive measurements, are among the least expensive on a per measurement basis. Passive measurements, which typically leave an electronic copy behind that can become the subject of a measurement study, often evolve into producing measurements of higher quality than those resulting from observations made by persons. However, attempts to gain Insights by conducting observations of one type or another will be used in instructional assessment for the foreseeable future. Due to this fact, methods of establishing the quality of observational measurements is of critical importance.

The characteristics of observational measurement systems such as active or passive collection when combined with knowledge about when measurements are to be collected, in-situation or post-situation, can be used to form a classification system that accommodates most observational measurement systems. The four unique combinations of these measurement system characteristics are: Active in-situation, Active Post-situation, Passive In-situation and Passive Post-situation. The classification of a measurement system is used to design a measurement system study and to select the proper analysis method for the resulting data.

For example, measuring behaviors as the behaviors occur by having a person complete a tally sheet or rating scale is a common example of an active in-situation measurement. When film or audio recordings of behaviors are collected and measured by a person, this is an example of an active post-situation measurement. Passive in-situation measurements are currently rare but could consist of measuring correlates of attention such as breathing, heart rate or stress indicators. An example of a Passive Post-situation measurement occurs when writing samples are scored using a text analyzer. These types of measurements are also currently rare. Passive Post-situation measurements are often coupled with automated storage of the measurements as well as automated first pass data analysis.

Regardless of the characteristics of a particular set of measurements, all observational measurements are the product of a measurement system that is comprised of the observational instrument(s) or device(s), the people and methods used to obtain measurements and the situations in which the measurements are collected. The total variation attributed to a measurement system designed for collecting observational measurements contains components related to the response, which is the variation we are trying to capture as well as components related to inaccuracy and the lack of objectivity, as shown in Equation 2.1. Accuracy indicates the centering of a measurement system with respect to samples for which the quantity of the response is known. These samples are known as standards. Objectivity indicates the amount of variation contained in each measurement attributable to the measurement system rather than to the response. Determining objectivity can involve standards or, under certain circumstances, be determined in a type of active in-situation study. The internationally accepted goals for the level of inaccuracy attributable to a measurement system is a shift of no more than 3% from the true value of each standard included in a study designed to detect inaccuracy. Regarding objectivity, the goal is to have the total amount of variation attributed to a lack of objectivity contained in each measurement of no more than 10% of the range of the measurements used in a study designed to detect a lack of objectivity

Equation 2.1: Components of Total Variance

$$\sigma^2_{Total} = \sigma^2_{Response} + \sigma^2_{Inaccuracy} + \sigma^2_{Objectivity}$$

2.2 Determining the % Inaccuracy

The types of standards used for determining the level of inaccuracy are developed through rigorous study of recorded or otherwise preserved instances of either the behaviors of interest or the products of the behaviors of interest. When the quantity of specific behaviors are of interest, the final values of the behaviors of interest contained in a particular standard are finalized using consensus techniques to resolve ambiguities. Accuracy and objectivity interact so that, if we envision thirty measurements landing somewhere in a target comprised of concentric rings, inaccuracy indicates the distance the average of the measurements is from the bull's-eye. The bulls-eye indicates the 'true' value of the behaviors or product of behaviors of interest contained in a particular standard. The level of objectivity is indicated by how dispersed the measurements are around the average of the measurements.

A study designed to determine the level of inaccuracy present in a measurement system is essentially a statistical demonstration that a measurement system is capable of measuring what is intended. For our purposes, inaccuracy will always be quantified as the percent inaccuracy resulting from direct comparison of observations to standards having known quantities of the behaviors or products of behaviors of interest, as calculated using Equation 2.2.

The goal of an accuracy study is to capture only shifts from each standard intrinsic to the measurement system without capturing variation related to environmental variables, personnel or other influences that together comprise a measurement systems lack of objectivity. When evaluating a measurement system, use a range of standards bracketing the range of the behaviors or products of behaviors likely to be encountered in the environment into which the measurement system will be implemented. This increases the usefulness of a measurement system because the behavior of a measurement system is only known within the range of standards included in the study.

Use of two standards allows a determination of the linear behavior of a measurement system. Using a third standard, midway between two standards representing the extremes, allows any nonlinear behavior of a measurement system to be detected. Regardless of the behaviors or products of behaviors of interest, developing standards that are useful for other accuracy studies as well as for studies designed to establish objectivity saves time and money. The % inaccuracy is reported separately for each standard where the organizations % inaccuracy requirement applies to each standard.

Equation 2.2: Percent Inaccuracy

$$\% \; Inaccuracy = \frac{\left|\overline{x}_{study} - Standard\right|}{Standard} \times 100$$

Conducting an inaccuracy study involves having a single person or device collect repeated measurements on each standard, completing the measurements for one standard before moving to the next. There should be enough repeated measurements to confidently estimate the mean for each standard, therefore twenty or thirty measurements are common unless collecting each measurement is expensive or difficult. If the measurement system being studied requires assembly, staging or calibration that could cause the measurements to vary, the measurement system should be completely prepared before measurement process begins. If calibration is required at set intervals, the measurements for all standards included in the accuracy study must be completed before recalibration is required. Variations introduced into measurements due to assembly, staging or calibration are captured and evaluated during an objectivity study.

Examples 2.1 and 2.2 illustrate most of the concepts encountered during studies designed to detect inaccuracy. Example 2.1 is for a measurement system that uses time intervals as the measurement. Example 2.2 is for a measurement system based around tracing a figure displayed on a computer screen. This second example of a measurement system could be easily adapted for use in a passive data collection system.

Example 2.1

In this example, the wait time between when a question is asked by an instructor and when the instructor selects someone to answer the question, if no one voluntarily responds, is the subject of an abbreviated accuracy study. The three standards for this accuracy study, A, B and C, consist of audio recordings of three wait time intervals. Each standard was measured five times by a single observer and the results are contained in Table 2.1.

Using the measurements collected for Sample A, the percent inaccuracy for this standard equals ((|10.4 - 10.2|) / 10.2) x 100 = 1.96% ~ 2.0%. Looking at the results for Standards B and C, notice that this measurement system complies with the goals for the %inaccuracy of 3% or less only for time intervals very near 10.2 seconds. Also notice that the %inaccuracy increases as the wait time interval becomes shorter. This trend is probably caused by the physiological limits of motor skill speed. In order to improve the range over which this measurement system is adequate, the timing method will need improvement. It should also be noted that 20 or 30 measurements should be collected for each standard even though only five measurements are contained in this example for brevity.

Table 2.1: Accuracy Study on Wait Times

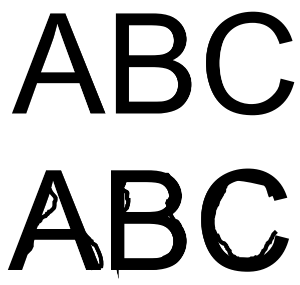
Standard	**A**	**B**	**C**
	10.3	8.4	5.9
	10.6	8.5	5.7
	10.1	8.3	5.8
	10.2	8.1	5.3
	10.9	8.7	5.5
Average	10.4	8.4	5.6
True Value	10.2	8.1	5.9
% Inaccuracy	**2**	**3.7**	**5.1**

Example 2.2

In this second example, an image consisting of three letters is traced by a student using a pen type mouse where the difference between the file size before and after the tracing becomes the measurement. The original image contains 9,906 bytes of information and the image with the student's tracing contains 19,896 bytes. So, the measurement equals a net gain of 9,990 bytes. In Figure 2.1, the original letters are displayed above the letters with tracing. For this measurement system, the difference in bytes is measured by the computer operating system and will remain the same regardless of the number of repeated measurements of these images. This means that the % inaccuracy for this measurement system is essentially 0%. This level of performance should be confirmed for different operating systems if more than one is to be used in other configurations of this measurement system.

This measurement system has an absolute 0 representing a perfect trace. Due to the fact that even slight departures from a perfect trace add hundreds of additional pixels, this measurement scale closely approximates a continuous ratio measurement scale in detail. This is important because most parametric forms of data analysis are only appropriate for measurements arising from a continuous scale i.e. having either an interval or a ratio level of detail.

Figure 2.1: Example of a Near Ratio Scale



2.3 Objectivity Overview

Now that issues related to accuracy studies have been introduced, the focus shifts to the topics required to understand and conduct an objectivity study. The total variation attributable to objectivity for a specific measurement system can be broken into two non-overlapping components known as repeatability and reproducibility as shown in Equation 2.3.

Equation 2.3: Components of Objectivity

$$\sigma^2_{Objectivity} = \sigma^2_{Repeatability} + \sigma^2_{Reproducibility}$$

Repeatability is the demonstrated ability of a measurement system to produce the same results, on the same occasion, for the same configuration and calibration.

Reproducibility is the demonstrated ability of a measurement system to produce the same results, on separate occasions after re-calibration, staging, etc., if applicable, of the same configuration.

It is only possible to determine both repeatability and reproducibility using variance components analysis on measurement systems that are configured for post-situation measurements. This is because some product of instruction is available which can be measured several times. Statistical independence of measurements is required for valid use of variance components analysis. This means that each measurement is influenced in no way by prior measurements. If the measurement system is configured for passive post-situation measurements, statistical independence of measurements is assured. One remote exception occurs when the measurement device contains adaptive programming that changes the sensitivity or centering of the device between measurements. Assured statistical independence is the gold standard for objectivity studies.

When measurements are collected active post-situation, the effects of previous exposure to each standard included in the study can be minimized by using techniques that aid forgetting or obscure the identity of the standards. Examples of these techniques are increasing the time interval between measurements collected by the same participant, including decoy standards mixed in with the actual standards and taking care to prevent communication between participants in the study regarding the study. If these methods are successful, dependable estimates of both the repeatability and reproducibility components of objectivity are obtainable. In practice, measurements collected actively post-situation constrain our ability to determine the reproducibility component of objectivity because statistical independence is always suspect to some degree.

When collecting measurements actively in-situation, identically reproducing separate instances of instruction or behaviors are not possible. The most common example of an active in-situation measurement occurs when an observer measures a representative sample of instruction, as the instruction occurs and produces a measurement based solely upon those observations. However, by having all perspective observers included in a measurement system, simultaneously measure each instance of instruction or behaviors, using identical methods without comparing the results during the measurement phase of the study, it becomes possible to estimate repeatability using simple linear regression. The linear regression technique for estimating the repeatability component of objectivity may also be the most practical method available for active post-situation measurements when statistical independence is suspect.

2.4 Objectivity of Active In-situation Measurements

Measurement systems based upon observations made by persons are typically organized around a collection of individual statements that are each followed by qualitative distinctions or categories indicating the degree to which the behaviors of interest are either present or absent. A coherent collection of these scales is known as an observational instrument. In most forms of observational instruments, an observer selects phrases, words or other descriptors from a continuum in order to produce a score or rating. Observational instruments based upon recording the frequency with which specific sets of behaviors occur in a preselected time interval are also common.

Each category for which tallies are collected or each statement and associated continuum is considered a scale and the observational instrument, as a whole, can be thought of as a summative scale. If resolution to the individual scale level is desired, the correlation between the scales comprising the summative scale for each possible pair combination of observers is calculated. One problem that arises when resolution to the individual scale level is desired is that the discrete nature of tallies or ratings causes the correlation between the observers ratings or tallies to be lower than that required for acceptable objectivity. Acceptable meaning that only 10% or less of the range of measurements resulting from a measurement system study is occupied by the probable location of a single measurement. Each measurement can be thought of as being bounded by a normal distribution placed symmetrically about each measurement. The normal distribution surrounding each measurement will be wider for a measurement system that has poor objectivity. The minimum number of measurement distribution widths that must fit end to end within the range of measurements contained in the measurement study equals 10.

If adequate correlation between the scale tallies or ratings between pairs of observers are below a certain threshold, the average tallies or ratings for sub-scales containing clusters of three or more individual scales are then correlated between each pair of observers. This often results in strong correlations as indicated by the size of the correlation coefficient. When this technique is used, the resolution of the measurement system is limited to the sub-scale level. More will be said about sub-scales when the procedures used in developing an observational instrument are discussed.

Equation 2.4: The Correlation Coefficient *r*

$$r = \frac{\sigma_{xy}^{2}}{\sqrt{\sigma_{x}^{2}\sigma_{y}^{2}}} = \frac{\dfrac{\sum (X-\bar{X})(Y-\bar{Y})}{n}}{\sqrt{\dfrac{\sum (X-\bar{X})^{2}}{n}}\sqrt{\dfrac{\sum (Y-\bar{Y})^{2}}{n}}}$$

When estimating the level of objectivity using simple linear regression, what is needed is a method of determining when the correlation coefficient between observers is large enough to meet the objectivity goal of 10% or less of the range of measurements attributed to a lack of objectivity. For this purpose we will use the Discrimination Index (D) shown in Equation 2.5. The Discrimination index indicates the number of distinct non-overlapping intervals along a linear regression line a measurement system can resolve. Figure 2.2 contains a geographic definition of D. In this figure, the maximum width of the ellipse, which occurs at the intersection of line projected to the regression line for the average of x and the average of y. The maximum ellipse width increases when a measurement system introduces a large amount of variation into the measurements. Because the ellipse width is larger, fewer ellipse widths will fit end to end within the length of the ellipse.

As can be seen in Equation 2.4, r equals the amount of the measurements produced by two observers vary together divided by the amount each varies individually. When measurements are plotted using the x-axis for one observer and the y-axis for another, r indicates the level of agreement between the two observers. In an active in-situation objectivity study or an active post-situation objectivity study not employing aids to forgetting and techniques that obscure the identity of the standards during the study, the value of D is calculated for all possible pairs of observers to be included in the final configuration of the measurement system. The smallest value of D for any pair of observers included in the final configuration of the measurement system is the number used to describe the objectivity of the measurement system as a whole.

When measurements which can only be positive numbers arising from a normal distribution are plotted on a regression plot, the measurements are said to have a bivariate normal distribution. In the general case, the correlation coefficient can range between 1 and -1 but for bivariate normal distributions where the minimum possible score equals zero, the correlation coefficient ranges between 0 and 1. Notice that when each X value is equivalent to a corresponding Y value, a special case that occurs where all points lie exactly on the regression line and so r equals 1.0 and D equals infinity.

For example, if D equals 5.0, only five widths around the regression line will fit end to end within the spread of the points along the regression line. The minimum number of distinct intervals resolved by a measurement system, under consideration for routine use, equals 10.0. When D is greater than or equal to 10.0, less than (1 / 10) x 100 = 10% of the range of measurements is occupied by the distribution surrounding each measurement. The correlation coefficient that results in D = 10.0 is r ~ 0.981.

Equation 2.5: The Discrimination Index D

$$D = \sqrt{\frac{1+r}{1-r}}$$

Strictly speaking, application of the correlation coefficient requires that the population distribution for the behaviors or products of behaviors being compared arise from a normally distributed population. The most obvious circumstance where measurements violate the normality assumption occurs when the range over which measurements occur is constrained so that the measurements bunch up at one end of the regression plot. A correctly designed observational instrument targets, on average, the midrange for the summative scale score while having individual scales or sub-scales that are sensitive to the range of behaviors or products of behaviors of interest. If only resolution to the sub-scale level is desired or possible, the average values for the sub-scales have the advantage of the central limit theorem regarding the assumption of bivariate normality.

Values of D corresponding to a range of values of r are contained in Table 2.2. For example, in Figure 2.2, r = 0.97. Referring to Table 2.2, D = 8.1 when r = 0.97. From the right two columns it is seen that a D value of less than 10 and greater than 5 receives a poor rating. Most of the issues concerning D are illustrated in Example 2.4

Figure 2.2: Geographic Definition of the Discrimination Index D

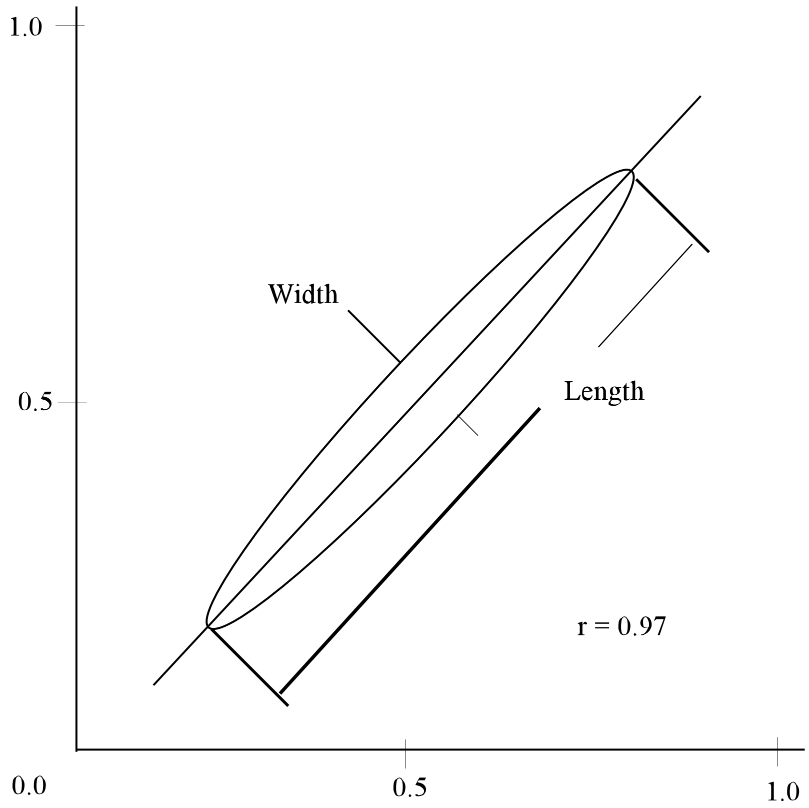


Table 2.2: D values Corresponding to r (*left*). Ratings for values of D (*right*).

r	D
0.80	3.0
0.90	4.4
0.91	4.6
0.92	4.9
0.92	5.0
0.93	5.3
0.94	5.7
0.96	7.0
0.97	8.1
0.98	9.9
0.981	10.2
0.990	14.1
0.995	20.0
0.998	31.6
0.999	44.7

D	Rating
3	Unacceptable
5	Poor
10	Acceptable
20	Good
50	Great
100	Excellent

Example 2.4

In the following example of an active in-situation objectivity study, five people were selected as observer candidates for a project designed to evaluate the quality of an organization’s instruction. The goal is to make recommendations either for or against implementation of this measurement system based upon direct observation of nine instances of instruction. Of the five perspective observers, only two observers and one alternate will be needed for operation of the measurement system if implemented. The alternate is to coordinate activities, attend meetings, perform analysis and generate reports when not observing. The observers will be selected by determining the three pairs of observers having the highest value of D.

During the objectivity study, all five perspective observers simultaneously observed and rated nine lectures using the same observational instrument for each lecture. The observers were physically separated during each lecture and were strongly encouraged not to discuss their results before completion of the objectivity study. The ratings for each observer candidate are shown at the top of Table 2.3. Below the ratings from each observer, the correlation coefficient between each pair of perspective observers is displayed in a form known as an inter-correlation matrix. Below the inter-correlation matrix are the discrimination index values corresponding to each correlation coefficient contained in the inter-correlation matrix. Only pairs of observers having values of $D \geq 10$ are of interest.

Table 2.3: Summative Ratings from Observers A thru E.

	A	**B**	**C**	**D**	**E**
1	36	31	26	21	34
2	41	21	29	25	35
3	39	26	32	39	25
4	73	76	80	81	22
5	85	101	115	111	38
6	101	133	129	136	62
7	115	154	134	151	94
8	131	177	165	170	121
9	199	187	193	183	142

Inter-correlation Matrix

	A	**B**	**C**	**D**	**E**
A	1.000	0.939	0.96	0.931	0.917
B		1.000	**0.988**	**0.994**	0.894
C			1.000	**0.991**	0.873
D				1.000	0.862
E					1.000

D Index

	A	**B**	**C**	**D**	**E**
A		5.7	7.0	5.3	4.8
B			**12.6**	**17.6**	4.2
C				**15.1**	3.8
D					3.7
E					

Upon reviewing Table 2.3, only combinations of observers B, C and D produce discrimination index values greater than 10.0. If these three observers are included in the final version of the measurement system, the discrimination index used to describe the objectivity of measurement system is 12.6 due to the size of the correlation between observers B and C. These configurations of this measurement system produces measurements having a percent of the range of measurements occupied by the variation attributable to a lack of the repeatability component of objectivity equaling (1 / 12.6) x 100 = 7.9%.

2.5 Objectivity of Post-situation Measurements

When determining the objectivity of a measurement system designed for post-situation use, if statistical independence of measurements is assured, it is possible to determine both the repeatability and reproducibility components of objectivity. This will frequently be the case for passive post-situation measurement systems. For measurement systems based upon observers rather than devices, aids to forgetting, decoy standards and other methods that obscure the identity of the standards must be used in order to approximate statistical independence in the study measurements. If assembly, staging or calibration variables are believed to be important, in order to capture these sources of variation, the equipment is set up before each individual participating in the study begins to measure the standards. Regardless of whether a device or a series of devices are identical or considered equivalent, a separate objectivity study is conducted for each configuration of the measurement system. One of the advantages accruing to post-situation objectivity studies is that the use of variance components analysis becomes possible.

In conducting a post-situation objectivity study, a preselected number of repeated measurements of each standard are collected on a preselected number of occasions. When a device or series of devices are used to perform the measurements, the occasions are typically, but not necessarily back to back. When persons collect the measurements there may be weeks or even months between subsequent measurements of the standards in order to allow forgetting to occur

The process of partitioning the variance begins by calculating the variance for repeatability and the variance for reproducibility for each standard. The variance for repeatability for a particular standard equals the average of the variances for each set of repeated measurements under each observer as shown in Equation 2.6. In Equations 2.6 thru 2.10, there are I = 1 to m observers or device operators, j = 1 to n standards and k repeated measurements collected by each observer or device operator on each standard.

Equation 2.6: Repeatability Variance for a Particular Standard

$$S_{r_j}^2 = \sum_{i=1}^{m} S_{ij}^2 \, / \, m$$

The reproducibility for a particular standard equals the variation calculated from the averages for each set of repeated measurements within a standard, as shown in Equation 2.7, minus a correction which amounts to the unexplained variation attributed to a single measurement as shown in Equation 2.8.

The amount of variation attributable to a single measurement equals the variance for repeatability for a particular standard divided by the number of measurements collected during each set of repeated measurements, k, by an observer or device. If the variance for reproducibility for a particular standard is less than the variation attributed to a single measurement, the variance for reproducibility will be negative for that particular standard and because negative variances are not allowed, the variance for reproducibility is set to 0.

Equation 2.7: Reproducibility Variance for a Particular Standard

$$S^2_{\bar{y}_j} = \frac{\sum_{i=1}^{m} (\bar{Y}_{ij} - \bar{\bar{Y}}_j)^2}{m-1}$$

Equation 2.8: Variation Attributable to a Single Measurement

$$S^2_{R_j} = S^2_{\bar{y}_j} - (S^2_{r_j} / k) \text{ or } 0 \text{ if } S^2_{R_j} < 0$$

The combined variance for repeatability and reproducibility equals the averages of the respective variances for all standards included in the study as shown in Equations 2.9 and 2.10. The standard deviation for repeatability or reproducibility for a measurement system equals the square root of the respective variance. The standard deviation for objectivity is calculated using Equation 2.11.

Equation 2.9: Variance and Standard Deviation Attributable to Repeatability

$$S^2_r = \frac{\sum_{j=1}^{n} S^2_{r_j}}{n} \text{ and } S_r = \sqrt{S^2_r}$$

Equation 2.10: Variance and Standard Deviation Attributable to Reproducibility

$$S^2_R = \frac{\sum_{j=1}^{n} S^2_{R_j}}{n} \text{ and } S_R = \sqrt{S^2_R}$$

In the analysis of variance method for conducting an objectivity study, each measurement can be thought of as having a symmetrically placed normal distribution surrounding the measurement. The width of this distribution equals 6 x S_r for repeatability, 6 x S_R for reproducibility and 6 x $S_{r\&R}$ for objectivity. The goal is to determine the number of the repeatability, reproducibility or objectivity distribution widths that will fit in a non-overlapping fashion, within the range of measurements included in the objectivity study. The number of distributions that will fit within the range of measurements included in the objectivity study is known as the number of distinct intervals.

The interpretation of the number of distinct intervals, a measurement system can resolve, is analogous to that for D even though these two methods are very different. In order to meet the goals for repeatability, reproducibility and objectivity, 10 or more distinct intervals must fit within the range of measurements.

Equation 2.11: Standard Deviation Attributable to both Repeatability and Reproducibility

$$S_{r\&R} = \sqrt{S_r^2 + S_R^2}$$

Either Equations 2.12a or 2.12b can be used to calculate the number of distinct intervals for either the repeatability or the reproducibility components of objectivity, as well as for objectivity as a whole. Six standard deviations are typically used as the width of the distribution surrounding a single measurement, but there are some that consider six standard deviations too conservative and so prefer Equation 2.12b. In either case, if rounding of a number of interval is desired, the number of interval is always rounded downward to the desired number of decimal points to insure that rounding produces a slightly conservative rating, as opposed to a slightly generous one.

Equation 2.12a: Distinct Intervals using 6 Standard Deviations

$$Distinct\ Intervals = \frac{Range}{6 \times S_r or\ S_R or\ S_{r\&R}}$$

Equation 2.12b: Distinct Intervals using 5.15 Standard Deviations

$$Distinct\ Intervals = \frac{Range}{5.15 \times S_r or\ S_R or\ S_{r\&R}}$$

The ratings for the analysis of variance method of determining repeatability, reproducibility and objectivity are obtained from Table 2.4. This table should be familiar because even though the number of distinct intervals is arrived at by a vastly different method from that used to calculate D, the fact remains that if a measurement system can theoretically resolve ten or more distinct intervals, the measurement system receives an acceptable rating.

Table 2.4: Objectivity Ratings for Number of Distinct Interval

Distinct Intervals	Rating
3	Unacceptable
5	Poor
10	Acceptable
20	Good
50	Great
100	Excellent

Example 2.5

In this example of a measurement system study, the wait time between when an instructor asks a question and when the instructor selects a student to answer the question, if no one volunteers, is the subject of an active post-situation objectivity study. In Example 2.1, observers measured the wait times for standards using a stop watch. Based upon that study, it was decided that stop watch timing lacked precision and so the recordings of the six wait time interval standards for this objectivity study were measured using sound editing software. Use of sound editing software allowed the wait time intervals to be measured to hundredths of a second. In this example, each of the n = 6 standards were measured k = 3 times by each of m = 2 observers. The measurements and preliminary calculations are contained in Table 2.5.

Table 2.5: Wait times and Preliminary Calculations

Sample	1	2	3	4	5	6	Avg. S	6 S
Obs. A	10.12	7.99	7.29	6.57	5.56	2.55		
	10.10	7.95	7.24	6.49	5.02	2.54		
	10.05	8.04	7.32	6.62	5.02	2.58		
Avg	10.09	7.99	7.28	6.56	5.20	2.56		
Var	0.0014	0.0019	0.0017	0.0046	0.0956	0.0005		
Obs. B	9.83	8.05	7.18	6.38	4.98	2.72		
	9.96	7.91	7.14	6.27	4.96	2.72		
	9.90	7.93	7.12	6.33	4.93	2.68		
Avg	9.90	7.96	7.14	6.33	4.96	2.71		
Var	0.0040	0.0055	0.0012	0.0027	0.0010	0.0005		
S_r	0.052	0.061	0.038	0.060	0.219	0.022	0.075	0.451
S_R	0.131	0.00[1]	0.097	0.159	0.113	0.105	0.101	0.605
$S_{r\&R}$							0.126	0.755

[1] The variance for reproducibility for sample 2 is negative so 0 is substituted

The calculations of Sr and SR for Standard 1are shown in Equations 2.13 and Equation 2.14

Equation 2.13: Standard Deviation Repeatability for Standard 1

$$S_{r1} = \sqrt{\frac{(0.00137 + 0.00401)}{2}} = 0.05185 \approx 0.052$$

Equation 2.14: Standard Deviation for Reproducibility for Standard 1

$$S_{R1} = \sqrt{0.01805 - \frac{0.05187}{3}} = 0.13097 \approx 0.131$$

The range of the measurements included in this study equals 10.12 - 2.54 = 7.58. The number of distinct intervals this measurement system can resolve for repeatability and reproducibility are shown in Equations 2.15 and 2.16.

Equation 2.15: Overall Repeatability Calculated using 6 Standard Deviations

$$Repeatability = \frac{7.58}{0.451} = 16.8$$

Equation 2.16: Overall Reproducibility Calculated using 6 Standard Deviations

$$Reproducibility = \frac{7.58}{0.605} = 12.5$$

For this measurement system, the number of non-overlapping intervals that can be resolved for repeatability equals 16.8 and for reproducibility equals 12.5, both of which are above the minimum of 10.0, and so this measurement receives acceptable rating for both components of objectivity. The number of distinct intervals this measurement system can resolve for objectivity equal 10.0 as shown in Equation 2.17. This active post-situation measurement system receives an acceptable rating for objectivity as well.

Equation 2.17: Overall Objectivity Calculated using 6 Standard Deviations

$$Objectivity = \frac{7.58}{0.755} = 10.0$$

2.6 Psychometric Development of an Observational Instrument

What follows is a brief discussion of certain aspects of the psychometric development process for creating an observational instrument. These calculations were carried out using a spreadsheet rather than dedicated software. There are many types of observational instruments, but as mentioned earlier, a common one involves counting the number of times a specific behavior occurs during a preselected interval such as a time interval or during a certain amount of instruction. Another common form of observational instrument, and the one for which an example of psychometric development is provided, involves a collection of scales which are used to detect the degree to which specific behaviors are either present or absence.

In this type of observational instrument, each statement is accompanied by a short list of descriptors in the form of adjectives, descriptive phrases or number analogs of adjectives or descriptive phrases representing different quantities of the sentiment described in the statement. The descriptors or number analogs are arranged in a continuum and when the statement and associated descriptors or number analogs are combined, the result is known as a scale. Each scale is the product of a scale specification and typically one or more scales are included, on the final form of an observational instrument, from each scale specification. Coherent groups of scales, typically from the same or very similar scale specifications, are known as sub-scales. The collection of all scales or sub-scales is known as a summative scale. The scale specifications and thereby the scales and sub-scales included in the observational instrument must accomplish the purpose for which the observational instrument is being developed.

At the outset, two or three times the number of scales to be included in the final form of the summative scale are generated and this collection of scales is known as an item pool. The item pool should contain two or more scales for each scale specification to insure that each scale specification survives the psychometric pruning process. Generating this rather large number of scales is necessary because the statistical performance of each individual scale, in relation to all other scales, cannot be predicted in advance.

After the item pool is complete, the focus of the development process begins by distilling the item pool into a summative scale that meets the purposes of the observational instrument in a way that maximizes the psychometric reliability of the resulting summative scale. When only resolution to the sub-scale level is possible, the psychometric reliability of each sub-scale is maximized. The psychometric reliability of a summative scale can be predicted for the desired number of sub-scales. This is accomplished by determining the minimum psychometric reliability each sub-scale must have in order to achieve the desired psychometric for the summative scale.

Psychometric reliability is a measure of the internal consistency of the scales and sub-scales that comprise a summative scale. Using modern methods, the psychometric reliability can be calculated from a single observational occasion if multiple observers measure the occasion simultaneously. It should be noted that obtaining a high level of psychometric reliability does not guarantee that a measurement system, based around an observational instrument, will succeed during an in-situation or post-situation objectivity study. This value does however provide the only concrete indication of likelihood of success available during the development process.

As mentioned earlier, scales can take several forms. A common form is known as a Behaviorally Anchored Rating Scale (BARS). In the BARS format, three to five declarative sentences or phrases describe a continuum concerning the presence or absence of specific behaviors. Examples of appropriate declarative sentences are: the subject takes additional responsibility the majority of times the opportunity arises or the subject takes additional responsibility infrequently.

Probably the most common scale format is known as a Likert scale. In a Likert scale, a statement describes a sentiment or an attitude and the observer selects from the continuum of descriptors or number analogs of descriptors in order to produce a score for each scale. The descriptors are of two types, the first of which are known as polar descriptors such as ‘Strongly Agree’ and ‘Strongly Disagree’. The polar descriptors form the ends of the continuum. The second type of descriptors are known as intermediate descriptors such as ‘Agree,’ or ‘Disagree’ that form the central region of the continuum.

In the basic format of the Likert scale, the intermediate descriptors are displayed as words between the two polar descriptors. The descriptors are then converted into number analogs for scoring purposes. One potential problem with displaying words in the continuum is that an observer may carry a predefined opinion or sentiment about certain descriptors and this can skew the observers’ choice. Another potential problem with displaying words, and then converting these to number analogs, is that this clerical task can introduce errors into the scoring.

A second format of the Likert scale, and one which largely avoids the issues associated with displaying words as descriptors, is called a graphical scale. In a graphical scale, only the polar descriptors are shown as words and scoring involves selecting a number analog from the continuum of number analogs located between the polar descriptors.

An example of a continuum for a graphic scale format is shown in Figure 2.3. A variation on this format is to remove the polar descriptors entirely and supply only abbreviations for descriptors as column headings while providing the number analogs in columns below each descriptor abbreviation. For example, in Figure 2.3, the continuum contains five gradations, so the column headings would be SA, A, N, D and SD and if a high score indicates more agreement with the statement, the columns would contain 5 for SA, 4 for A, 3 for N, 2 for D and 1 for SD.

Figure 2.3: Example of a Graphical Scale form of a Likert Scale

Strongly Agree 1 2 3 4 5 Strongly Disagree

One serious problem that can arise when using Likert scales is known as a response set. A response set is a tendency to respond in a certain and often idiosyncratic pattern, to a particular scale format. A tendency to agree with all statements regardless of content, or to select the same number analog for all scales are two common response sets. One way to counteract response sets is to take extra time and ensure that the statements adhere to a few guidelines. Specifically, all statements should be short, in present tense and avoid universals, indefinite qualifiers and negatives. The vocabulary and sentence structure should be simple, appropriate and avoid content that is factual or has more than one interpretation. Also, avoid statements that are likely to be advocated by most people or by almost no one.

Using a mix of positive statements and negative statements has the effect of forcing the observer to read each statement carefully before selecting a descriptor or number analog. A positively worded statement is converted into a negatively worded statement by changing a few words in a way that reverses the meaning of the polar descriptors. The definiteness of the score produced by each scale is improved by using an even number of descriptors or number analogs containing no neutral option. This has the effect of forcing the observer to express an opinion either in favor of or against the sentiment expressed in the statement. In a sub-scale or summative scale comprised of Likert scales of both the positively and negatively worded variety, the score for each negatively worded scale must first be converted into the analogous score that would result if the statement were positively worded before proceeding with scoring. For example, a rating of 2 on a 1 to 8 scale, having a negatively worded statement, would be converted to the analogous value for a positively worded statement which equals 7.

One way to avoid the need to convert scale scores for sub-scales or a summative scale having a negatively worded statements is to reverse the order of the intermediate descriptors or number analogs accompanying a negative statement. This has the same effect of placing the scores in the correct order without having to reorder the scores in a separate step. The number of descriptors or number analogs contained in a Likert scale that maximizes the psychometric reliability of the resulting sub-scales or summative scale has been studied extensively. The consensus of these studies indicates that the psychometric reliability increases rapidly as the number of descriptors or number analogs contained in each scale increase up to around 7 or 8. The benefit from adding additional descriptors or number analogs continues to increase but at a decreasing rate up to around 11. The reason for this is that, from a statistical standpoint, the larger the number of descriptors or number analogs, the closer a scale approaches a continuous scale of measurement. As mentioned earlier, discrete measurements tend to inflate the variance in formulas that are designed for continuous measurements. Unfortunately, even with 11 gradations in the continuum, the minimum distance between any two gradations is still large when compared to a truly continuous scale of measurement and this has the effect of limiting the maximum obtainable level of psychometric reliability of the resulting sub-scale or summative scale.

Now that a few of the considerations for constructing an item pool have been presented, our discussion turns to pruning the item pool in a way that maximizes the psychometric reliability of both the sub-scales and the summative scale.

The calculations used to select the set of individual scales from the item pool that maximizes the psychometric reliability are almost identical to that used for developing a standardized test, a topic addressed in Chapter 3. In fact, if the scores for each individual scale are converted to dichotomous scores of 1 or 0 based upon whether the scale score is above or below the midpoint of the continuum, the same equation is used to calculate the psychometric reliability of a sub-scale or summative scale as that used for a standardized test.

If the individual scales are believed to contain information that warrants interpretation beyond a dichotomy, the most widely used measure of psychometric reliability is Coefficient Alpha ( ). Even though the same character is used to describe the probability of committing a Type I error, these topics are not to be confused. The formula for coefficient alpha is shown in Equation 2.18.

Equation 2.18: Coefficient Alpha

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma^2}\right)$$

After an unpruned summative scale has been assembled, each of the scales is used by the observers similar to those to be included in the proposed measurement system. Including a few extra observers in the study is advisable. The observers completely score the instance of instruction before any attempt is made to estimate the statistical performance of each scale in relation to all other scales. An observation session can involve active post-situation observations of standards or be accomplished by simultaneous active in-situation observations.

The endpoint of the pruning process is reached when the remaining summative scale or sub-scale contains one or more representatives from each scale specification and the summative scale or sub-scale meets the goals for psychometric reliability. A summative scale or sub-scale that is comprised of scales that are highly correlated with the sum of the scale scores maximizes the psychometric reliability. In addition to having individual scales that are highly correlated with the sum of the scale scores, having scales that are moderately correlated with the other individual scales included in a sub-scale or summative scale also maximizes psychometric reliability.

Specifically, in Equation 2.18, the quotient formed by the sum of the item variances divided by the variance of the scale scores is minimized when the sum of the item variance is small in relation to the variance of the scale scores. The variance of an individual scale is small if all of the scores produced by all observers are the same or nearly the same. When quotient this is small and subtracted from one, the result is a high value of . Hypothetically, if this quotient equals 0 and k is very large k / (k - 1) x (1 - ~ 0) = ~ 1. Values of > 0.95 for a summative scale are common in actual practice.

When resolution to the individual scale level is desirable, the goal then becomes to maximize the psychometric reliability of each sub-scale and through this mechanism to maximize the psychometric reliability of the summative scale. Psychometric development of one sub-scale consisting of five Likert scales, having seven gradations in the continuum, is presented next as a way of introducing and explaining the details of the psychometric pruning process.

Example 2.6

In this example, each scale has two polar descriptors and the numbers 1 thru 7 represent gradations of the sentiment expressed in the statement. In Table 2.6, each triplet of scales represent a single scale specification. The final sub-scale must contain one scale from each of the five scale specifications. If possible, it is desirable that the summative scale and the resultant measurement system have resolution to the individual scale level. The number of sub-scales of equivalent quality, to the one being developed, that are required to achieve a certain level of psychometric reliability for the summative scale can be estimated.

As mentioned earlier, it is impossible to estimate the statistical performance of an individual scale in relation to all other scales from the outset, so having a sufficient number of scales representing each specification is essential. Table 2.6 contains the scale scores resulting from observations collected from 14 individuals simultaneously scoring a single standard using a sub-scale containing 15 scales. The scores for scales that were negatively worded have been converted to the number analogs that would have resulted for scales having positively worded statements. In Table 2.6, the bottom row contains the variance of the scores within each scale over all observers. The sum of these item variances equals 15.02 and the variance of the scale scores equals 29.96. These two values will be used shortly in calculating .

Table 2.6: Scale Scores for the Initial Sub-scale for all 15 Observers

Obs.	1a	1b	1c	2a	2b	2c	3a	3b	3c	4a	4b	4c	5a	5b	5c		Tot	
1	4	2	4	5	3	6	3	3	5	5	4	4	5	4	4		61	
2	3	2	4	3	4	4	4	2	3	3	3	4	4	2	3		48	
3	2	2	2	4	4	2	2	2	2	4	4	5	3	2	4		44	
4	3	4	2	4	4	3	4	4	4	4	3	2	3	3	3		50	
5	6	5	4	3	4	5	4	4	5	3	3	3	4	4	2		59	
6	4	4	3	3	3	3	1	4	4	3	4	2	4	4	4		50	
7	4	4	4	5	3	3	3	6	3	5	5	3	4	4	3		59	
8	3	3	4	3	3	2	4	3	5	3	4	4	3	3	3		50	
9	4	3	2	4	3	2	3	4	2	4	2	4	2	4	4		47	
10	3	4	3	4	2	5	4	3	4	4	3	4	2	3	2		50	
11	4	3	4	2	4	2	5	2	4	2	3	4	3	2	2		46	
12	2	1	4	3	3	4	1	2	3	3	4	6	4	2	4		46	
13	4	4	3	4	5	3	2	3	4	4	4	1	3	3	3		50	
14	3	4	4	5	5	4	1	4	3	5	4	3	5	6	4	∑**Var i**	60	
Var i	0.96	1.17	0.66	0.78	0.67	1.53	1.64	1.20	0.94	0.78	0.53	1.54	0.82	1.20	0.60	15.02	30	**S^2**

Table 2.7 contains the inter-correlation matrix for all combinations of scales. Attention should also be paid to scales that are highly or perfectly correlated with other scales. This is because highly or perfectly correlated scales essentially measure the same aspect of behavior, and because of this, do not provide information beyond that provided by including one scale or the other. Perfectly correlated scales will hopefully be eliminated during the pruning process, but regardless, perfectly correlated scales should not be included in the final form of a sub-scale or summative scale. Upon inspection of this table, it can be seen that scales 2a and 4a are perfectly correlated. If these two scales should survive the pruning process, either scale 2a or 4a should be eliminated from the final form of the sub-scale.

Table 2.7: Inter-correlation Matrix for Initial Sub-scale

	1b	1c	2a	2b	2c	3a	3b	3c	4a	4b	4c	5a	5b	5c
1a	0.64	0.22	-0.08	0.09	0.24	0.31	0.46	0.49	-0.08	-0.20	-0.50	0.12	0.40	-0.42
1b		-0.09	0.14	0.18	0.04	0.17	0.67	0.34	0.14	-0.06	-0.77	-0.11	0.55	-0.48
1c			-0.16	0.02	0.42	0.09	-0.03	0.43	-0.16	0.38	0.18	0.63	0.13	-0.24
2a				0.03	0.31	-0.34	0.53	-0.20	**1.00**	0.37	-0.20	0.27	0.60	0.40
2b					-0.17	-0.17	-0.10	-0.10	0.03	0.05	-0.42	0.29	0.14	0.03
2c						-0.03	0.02	0.42	0.31	0.05	0.05	0.51	0.28	-0.10
3a							-0.09	0.38	-0.34	-0.49	0.02	-0.40	-0.34	-0.78
3b								0.03	0.53	0.24	-0.53	0.14	0.70	0.01
3c									-0.20	0.09	-0.33	0.20	0.10	-0.47
4a										0.37	-0.20	0.27	0.60	0.40
4b											-0.08	0.54	0.15	0.29
4c												-0.03	-0.42	0.19
5a													0.43	0.36
5b														0.26

Table 2.8 contains the total scores for each scale once the scores located in the analogous column in Table 2.6 has been removed. The score of the scale being correlated is removed because the scale score is auto-correlated with the total scores and this can inflate the correlation. For example, the total score for Observer 1, for all 15 scales, was 61 as shown in the rightmost column of Table 2.6. Once the score for scale 1a is removed, the total score for Observer 1 equals 61 - 4 = 57, as bolded in the first data row and column in Table 2.8.

Using Equation 2.18, coefficient alpha equals 0.53 for the fifteen scales contained in Table 2.6 as shown in Equation 2.19. These values are located in the first data row on Table 2.8.1 in the row labeled ‘none’ which indicates that no scales have been removed before calculating .

Equation 2.19: for Calculated from All Scales

$$\alpha = \frac{15}{14}\left(1 - \frac{15.02}{29.96}\right) = 0.53$$

Eliminating the most advantageous scale during each iteration of the pruning process involves calculating coefficient alpha for the sub-scale that would result from the removal of each of the remaining scales, one scale at a time. The scale that ‘if removed’ maximizes the size of coefficient alpha is permanently removed during each iteration. Notice that in the last row of Table 2.8, the correlation coefficients for scales 2b, 3a, 4c and 5c are negative. Generally, scales that are negatively or weakly correlated with the total scores, at each iteration, tend to be eliminated first.

Table 2.8: Total Scores Once the Score for the Scale Listed in the First Row is Removed

	1a	**1b**	**1c**	**2a**	**2b**	**2c**	**3a**	**3b**	**3c**	**4a**	**4b**	**4c**	**5a**	**5b**	**5c**
1	**57**	59	57	56	58	55	58	58	56	56	57	57	56	57	57
2	45	46	44	45	44	44	44	46	45	45	45	44	44	46	45
3	42	42	42	40	40	42	42	42	42	40	40	39	41	42	40
4	47	46	48	46	46	47	46	46	46	46	47	48	47	47	47
5	53	54	55	56	55	54	55	55	54	56	56	56	55	55	57
6	46	46	47	47	47	47	49	46	46	47	46	48	46	46	46
7	55	55	55	54	56	56	56	53	56	54	54	56	55	55	56
8	47	47	46	47	47	48	46	47	45	47	46	46	47	47	47
9	43	44	45	43	44	45	44	43	45	43	45	43	45	43	43
10	47	46	47	46	48	45	46	47	46	46	47	46	48	47	48
11	42	43	42	44	42	44	41	44	42	44	43	42	43	44	44
12	44	45	42	43	43	42	45	44	43	43	42	40	42	44	42
13	46	46	47	46	45	47	48	47	46	46	46	49	47	47	47
14	57	56	56	55	55	56	59	56	57	55	56	57	55	54	56
r	0.38	0.27	0.36	0.47	-0.06	0.45	-0.27	0.48	0.27	0.47	0.24	-0.51	0.56	0.67	-0.16

In Table 2.8.1, the right most column contains the coefficient alpha values that would result if the scale listed in the leftmost column is removed before calculating coefficient alpha. The scale removed in a particular row is restored before removing the scale indicated in the next row. These calculations are repeated row by row until all surviving scales have been removed and coefficient alpha calculated. These ‘if removed’ calculations indicate that during the first iteration of the pruning process, coefficient alpha is maximized when scale 4c is removed.

Table 2.8.1: ‘If Removed’ Iteration 1

Removed	S_i	S	k	
none	15.02	29.96	15	0.534
1a	14.06	23.16	14	0.423
1b	13.85	27.25	14	0.530
1c	14.36	24.24	14	0.439
2a	14.24	27.20	14	0.513
2b	14.35	31.25	14	0.582
2c	13.49	22.01	14	0.417
3a	13.38	29.04	14	0.581
3b	13.82	24.81	14	0.477
3c	14.08	23.89	14	0.442
4a	14.24	27.20	14	0.513
4b	14.49	27.21	14	0.503
4c	**13.48**	**33.81**	**14**	**0.648**
5a	14.20	24.04	14	0.441
5b	13.82	24.85	14	0.478
5c	14.42	32.64	14	0.601

The ‘if removed’ values of coefficient alpha are recalculated with scale 4c removed as shown in Table 2.8.2. The next scale which ‘if removed’ maximizes the size of coefficient alpha is scale 5c which results in a coefficient alpha value of 0.703.

Table 2.8.2: ‘If Removed’ Iteration 2

Removed	S_i	S	k	
1a	12.52	26.21	13	0.566
1b	12.31	29.80	13	0.636
1c	12.82	28.09	13	0.589
2a	12.70	31.25	13	0.643
2b	12.81	35.00	13	0.687
2c	11.95	25.76	13	0.581
3a	11.84	33.09	13	0.696
3b	12.28	27.36	13	0.597
3c	12.54	27.04	13	0.581
4a	12.70	31.25	13	0.643
4b	12.95	30.96	13	0.630
5a	12.66	27.49	13	0.584
5b	12.28	28.00	13	0.608
5c	**12.88**	**36.69**	**13**	**0.703**

Once scale 5c is eliminated, coefficient alpha is maximized by eliminating scale 2b as shown in Table 2.8.3.

Table 2.8.3: 'If Removed' Iteration 3

Removed	**S_i**	**S**	**k**	
1a	11.92	28.65	12	0.637
1b	11.71	31.76	12	0.689
1c	12.22	30.49	12	0.654
2a	12.10	34.41	12	0.707
2b	**12.21**	**37.96**	**12**	**0.740**
2c	11.35	27.84	12	0.646
3a	11.24	34.89	12	0.739
3b	11.68	30.04	12	0.667
3c	11.94	29.24	12	0.645
4a	12.10	34.41	12	0.707
4b	12.35	34.04	12	0.695
5a	12.06	30.61	12	0.661
5b	11.68	30.96	12	0.679

Next, coefficient alpha is maximized by eliminating scale 3a, as shown in Table 2.8.4.

Once scales 4c, 5c, 2b and 3a are removed, all of the scales that were negatively or weakly correlated with the total scores in Table 2.6 have been eliminated. Next our 'if removed' analysis indicates that removing either scale 2a or 4a, which as you'll recall from the inter-correlation matrix are perfectly correlated, will maximize the value of coefficient alpha. Scale 2a was selected for elimination during this iteration.

Table 2.8.4: 'If Removed' Iteration 4

Removed	**S_i**	**S**	**k**	
1a	11.25	29.96	11	0.687
1b	11.04	32.85	11	0.730
1c	11.55	31.64	11	0.698
2a	11.43	35.40	11	0.745
2c	10.68	28.81	11	0.692
3a	**10.57**	**36.24**	**11**	**0.779**
3b	11.01	31.01	11	0.709
3c	11.27	30.29	11	0.691
4a	11.43	35.40	11	0.745
4b	11.68	35.21	11	0.735
5a	11.39	32.24	11	0.711
5b	11.01	31.85	11	0.720

Table 2.8.5: ‘If Removed’ Iteration 5

Removed	S_i	S	k	
1a	9.61	28.40	10	0.735
1b	9.40	31.41	10	0.779
1c	9.91	30.44	10	0.749
2a	**9.79**	**33.56**	**10**	**0.787**
2c	9.04	27.89	10	0.751
3b	9.37	29.09	10	0.753
3c	9.63	29.29	10	0.746
4a	9.79	33.56	10	0.787
4b	10.04	32.89	10	0.772
5a	9.75	30.16	10	0.752
5b	9.37	29.81	10	0.762

Next, coefficient alpha is maximized when scale 4a is removed. It would not have mattered whether scale 4a was eliminated before scale 2a during iteration 5. Both of these scales were going to fall out during iterations 5 and 6.

Table 2.8.6: ‘If Removed’ Iteration 6

Removed	S_i	S	k	
1a	8.83	25.56	9	0.736
1b	8.62	28.45	9	0.784
1c	9.13	27.64	9	0.753
2c	8.26	25.61	9	0.762
3b	8.59	27.01	9	0.767
3c	8.85	26.29	9	0.746
4a	**9.01**	**32.00**	**9**	**0.808**
4b	9.26	30.61	9	0.785
5a	8.97	27.64	9	0.760
5b	8.59	27.45	9	0.773

Once scales 2a and 4a are removed, the next scale indicated is scale 4b. This is unfortunate because if this scale is removed, all scales representing scale specification 4 will have been removed. So, scale 4b is left and the next best ‘if removed’ scale is selected which is scale 1b as shown in Table 2.8.7.

Table 2.8.7: ‘If Removed’ Iteration 7

Removed	S_i	S	k	
1a	8.05	23.84	8	0.757
1b	**7.84**	**26.61**	**8**	**0.806**
1c	8.35	25.96	8	0.775
2c	7.48	24.45	8	0.793
3b	7.81	26.05	8	0.800
3c	8.07	24.41	8	0.765
4b	8.48	29.45	8	0.814
5a	8.19	26.24	8	0.786
5b	7.81	26.21	8	0.802

In iterations 8, 9 and 10, disregarding the benefits of removing 4b, scales 3b, 5b and 1a are removed next, in that respective order as shown in Tables 2.8.8, 2.8.9 and 2.8.10. Once scale 1a is removed, the final sub-scale consists of scale's 1c, 2c, 3c, 4b and 5a. The coefficient alpha value for this sub-scale equals 0.850.

Table 2.8.8: 'If Removed' Iteration 8

Removed	S_i	S	k	a
1a	6.88	19.69	7	0.759
1c	7.18	20.65	7	0.761
2c	6.31	19.36	7	0.786
3b	**6.64**	**22.16**	**7**	**0.817**
3c	6.90	19.80	7	0.760
4b	7.31	23.96	7	0.811
5a	7.02	20.61	7	0.769
5b	6.64	21.64	7	0.809

Table 2.8.9: 'If Removed' Iteration 9

Removed	S_i	S	k	
1a	5.68	16.44	6	0.785
1c	5.98	16.40	6	0.762
2c	5.11	14.61	6	0.780
3c	5.70	15.45	6	0.757
4b	6.11	20.01	6	0.834
5a	5.82	16.36	6	0.773
5b	**5.44**	**18.49**	**6**	**0.847**

Table 2.8.10: 'If Removed' Iteration 10

Removed	Si	S	k	
1a	**4.48**	**14.01**	**5**	**0.85**
1c	4.78	13.01	5	0.791
2c	3.91	11.44	5	0.823
3c	4.50	12.36	5	0.795
4b	4.91	16.44	5	0.877
5a	4.62	13.05	5	0.807

Ideally, the final form of the sub-scale should not contain any scales that are not statistically significantly correlated with the total scores for the sub-scale. By substituting a few values into Equation 2.20, it is possible to determine the minimum value of r required at different levels of significance. So, for n = 14, r = 0.453 is required for significance at the 0.05 level and r = 0.362 is required at the 0.10 level of significance.

The correlation between scale 4b and the total scores, which equals 0.31, is not significant at the 0.10 level. The scales representing this scale specification have been problematic throughout the analysis due to the fact that at iteration 7, this scale specification would have been eliminated from the sub-scale. This points to a need to repair one or more scales representing scale specification 4 and include these reworked scales during pilot testing or as the subject of another measurement study of the sub-scale.

Table 2.9: Correlation of Scales to the Sums of Scale Scores at Each Iteration

	1a	**1b**	**1c**	**2a**	**2b**	**2c**	**3a**	**3b**	**3c**	**4a**	**4b**	**4c**	**5a**	**5b**	**5c**
0	0.38	0.27	0.36	0.47	-0.06	0.45	-0.27	0.48	0.27	0.47	0.24	-0.51	0.56	0.67	-0.16
1	0.46	0.43	0.29	0.47	0.03	0.39	-0.26	0.56	0.31	0.47	0.24		0.51	0.71	-0.18
2	0.51	0.49	0.32	0.41	0.03	0.39	-0.16	0.55	0.37	0.41	0.19		0.45	0.65	
3	0.50	0.47	0.32	0.41		0.43	-0.14	0.57	0.40	0.41	0.19		0.41	0.64	
4	0.42	0.42	0.30	0.48		0.43		0.59	0.30	0.48	0.30		0.50	0.72	
5	0.48	0.44	0.35			0.42		0.55	0.37	0.35	0.26		0.50	0.68	
6	0.54	0.45	0.41			0.39		0.48	0.44		0.22		0.49	0.61	
7	0.44		0.49			0.44		0.37	0.41		0.26		0.60	0.54	
8	0.36		0.58			0.51			0.46		0.23		0.65	0.39	
9	0.28		0.65			0.51			0.52		0.21		0.60		
10			0.66			0.50			0.41		0.31		0.66		

Equation 2.20: *r* Required for Significance

$$r = \frac{t_{n-2,\alpha}}{\sqrt{(n-2)+t_{n-2,\alpha}^{2}}}$$

One final question remains unanswered and that being, what would the value of coefficient alpha be for a summative scale if the summative scale consists of four sub-scales of equivalent quality to the sub-scale just developed. In Equation 2.21, K equals the expected number of scales comprising the summative scale, which equals 20, divided by the number of scales expected to be in each sub-scale which equals 5, so K = 20 / 5 = 4 sub-scales. Using Equation 2.21, the resulting value of coefficient alpha value should be approximately = 0.96 if all sub-scales are of equivalent quality, as shown in Equation 2.22. This coefficient alpha value is in the range that is typical for a summative scale comprised of Likert scales.

Equation 2.21: Predicted Coefficient Alpha for K Sub-scales

$$\hat{\alpha}_{total\ scale} = \frac{K \times r_{subscale}}{1+(K-1)r_{subscale}}$$

$$where,$$

$$K = \frac{total\ expected\ number\ of\ scales}{number\ of\ scales\ in\ subscale}$$

Equation 2.22: Predicted Coefficient Alpha for 4 Sub-scales Having = 0.85

$$\hat{\alpha}_{total\ scale} = \frac{4\times 0.85}{1+(4-1)\times 0.85} = 0.9577$$

2.7 Objectivity of a Likert Type Summative Scale

Maximizing the psychometric reliability of each sub-scale results in a summative scale having high internal consistency. Internal consistency is desirable because this quality is predictive of objectivity. Regardless of the size of coefficient alpha, before wide scale implementation, an objectivity study is conducted to determine the objectivity of the measurement system. Determining only the repeatability component of objectivity is possible if statistical independence of measurements is not assured.

For the sub-scale just developed, an active post-situation objectivity study employing aids to forgetting and analysis of variance techniques was conducted. This allows estimation of both the repeatability and reproducibility components of objectivity. The study included three standards which span the expected range of the behaviors likely to be encountered upon implementation. Having three standards also allows detection of nonlinear behavior in the measurement system. In Example 2.7, the sub-scale developed studied three observers on two separate occasions employing aids to forgetting.

Example 2.7

In this example, three observers A, B and C score three video recordings containing different quantities of the behaviors or characteristics of interest on two separate occasions using the sub-scale developed earlier. The occasions were separated by one month and each observation session consisted of six video recordings randomized before each session, of which, only three recordings are actually standards for the objectivity study. Notice that because the sub-scale scores are used in the analysis, resolution to the individual scale level is implied. Table 2.10 contains the measurements and calculations for the complete post-situation objectivity study.

Table 2.10: Repeatability and Reproducibility study

		Standard						
Obs.	**Ses.**	**1**	**2**	**3**				
A	1	8	21	31				
	2	6	23	33				
B	1	9	23	29				
	2	5	24	34				
C	1	9	22	29				
	2	7	25	31				
avg.	**A**	7.0	22.0	32.0				
avg.	**B**	7.0	23.5	31.5				
avg.	**C**	8.0	23.5	30.0				
S^2 avg.		0.22	0.50	0.72				
Var.	**A**	1.00	1.00	1.00				
Var.	**B**	4.00	0.25	6.25				
Var.	**C**	1.00	2.25	1.00				
S^2 var.		2.00	1.17	2.75				
S^2_r		0.073	0.167	0.240	**S**	**5.15S**	**Range**	**Obj.**
S_r		0.270	0.409	0.490	0.390	2.009	29	**14.4**
S^2_R		0.180	0.420	0.600				
S_R		0.424	0.648	0.775	0.616	3.172	29	**9.1**
S_{r+R}					0.729	3.755	29	**7.7**

By referring to bottom right portion of Table 2.10, the repeatability component of objectivity indicates that there are 14.4 distinct intervals between the minimum and maximum values included in the study, which equals 34 - 5 = 29. This measurement system obtains a rating approximately midway between acceptable and good for the repeatability component of objectivity. This measurement system can resolve 9.14 distinct intervals for the reproducibility component of objectivity which is just slightly below an acceptable rating. Notice that the more generous method for calculating the number of distinct intervals, of dividing the range by 5.15 standard deviations as opposed to 6.0 standard deviations, were used in this analysis. When the two components of objectivity are combined using Equation 2.13, the overall number of distinct intervals resolved by this measurement system equals 7.72 which results in a rating midway between poor and acceptable.

This is problematic because this places limits on the interpretation of measurements occurring over a series of occasions after implementation. It may be possible to improve the reproducibility by having three intact teams containing two or three observers score each standard included the study. In this configuration of an objectivity study, the average rating for each team becomes the measurement. Of course, only intact teams will be able to perform the observations simultaneously upon implementation.

Another strategy for improving the objectivity of this sub-scale is to add one or more scales to the sub-scale. In most psychometric development efforts, increasing the number of scales improves both components of objectivity. Also, using scales containing 10 or 11 gradations in the continuum associated with each statement of each Likert scale would likely increase the range of the measurements. If increasing the range of measurements, by increasing the number of gradations, does not result in a proportional increase of the variation between observers, both the repeatability and reproducibility components of objectivity will improve.

2.8 Concluding Remarks

The underlying conditions for the most successful measurement systems are that each measurement contains less than 3% inaccuracy, to all known causes, and have a lack of objectivity of 10% or less of the range of measurements likely to be encountered in the environment into which the measurement system is to be implemented. Passive post-situation measurement techniques, especially those relying upon software products, typically meet these criteria where as measurements collected by persons are typically below these criteria if resolution to the scale level is desired .

Due to the fact that cost containment is an integral aspect of a quality control program, the goal is to move from infrequent and costly measurements toward more frequent inexpensive measurements. Nowhere are the relative costs of measurements more evident than between active measurement systems and passive measurement systems. In order to move from instructional processes that are configured to instruct groups of students to ones that are configured to instruct individuals, large quantities of low cost high quality measurements will be necessary. It is difficult to imagine this happening in the absence of passive data collection techniques used to evaluate the products of behaviors rather than observations of behaviors directly. A great challenge in both the fields of education and training is to develop passive in-situation and passive post-situation measurement systems.

3

Tests and Examinations as Measurements

3.1 Tests and Examinations Defined

As mentioned in Chapter 1, measurements are the product of a measurement system comprised of the instruments, people and methods used in the collection of each measurement. The most common form of measurement system for use in determining the efficacy of instruction is a collection of questions known as either a test or an examination. In a test or examination, students supply answers to questions where the number of questions answered correctly provides the measurement. For the purposes of this text, when the collection of questions are selected using psychometric methods, the assessment is known as a standardized test or a test. When the questions are selected by other methods, such as by an instructor, the assessment is known as an instructor generated examination or simply an examination. The value of a measurement depends upon the amount of information the measurement contains that can be extracted and used to either maintain the status quo or improve upon the status quo of the instruction offered by an organization. The focus of this chapter is to illustrate strategies for developing tests and examinations that, when combined with thoughtfully configured measurement systems, produce measurements possessing the maximal amount of extractable information.

The total variation attributed to these types of measurement systems contains components related to knowledge, accuracy and objectivity as shown in Equation 3.1. A perfect measurement system would produce measurements that contain only variation attributable to differences in knowledge where the variation attributable to a lack of accuracy and objectivity equals zero. In Chapter 2, accuracy is expressed as the percent inaccuracy contained in a measurement and the %inaccuracy is determined by a statistical demonstration that a measurement system is capable of measuring what is intended. This same definition is used when discussing tests and examinations. However, in the case of tests and examinations, the %inaccuracy is comprised of components that are minimized individually. Objectivity indicates the likelihood that a set of measurements can be duplicated under the same or similar measurement conditions. Estimating the objectivity of either a test or an examination is more difficult than for an observational measurement system because students are irreversibly altered by exposure to the examination or test. There are strategies available for determining objectivity of measurements systems based around an examination or test and these are discussed in detail in this chapter. Our discussion begins with the topics related to the %inaccuracy of these types of measurements.

Equation 3.1: Total Variance Components

$$\sigma^2_{Total} = \sigma^2_{Knowledge} + \sigma^2_{Accuracy} + \sigma^2_{Objectivity}$$

3.2 Accuracy of Examinations and Tests

As is the case for observational measurement systems, the goal for the %inaccuracy in this chapter is consistent with the internationally accepted limits of 3% or less. Standards are used to determine the accuracy of measurement systems based upon observations where, the difference between the actual measurements and the true quantity of the behaviors or characteristics of interest equals the amount of inaccuracy attributable to the measurement system. For a measurement system based upon a test or examination, the %inaccuracy is determined by application of both logical and statistical proofs. Specifically, producing accurate measurements using a test or examination involves minimizing three sources of inaccuracy: content sampling errors, misclassification errors and the influence of guessing as shown in Equation 3.2. These sources of inaccuracy occur to differing degrees and for different reasons in standardized tests as opposed to instructor-generated examinations. Where differences occur, these will be illustrated for each type of assessment separately.

Equation 3.2: Inaccuracy Variance Component

$$\sigma^2_{Inaccuracy} = \sigma^2_{Content\ Sampling} + \sigma^2_{Misclassification} + \sigma^2_{Guessing}$$

3.3 Content Sampling Errors

Concerning standardized tests, content sampling is believed to be the largest source of inaccuracy. This is because standardized tests are typically comprised of a relatively small sample of questions selected from a large body of knowledge known as a domain. This conception is formally known as the Domain Sampling Model. In order to insure adequate and uniform sampling of a domain, question specifications are generated and this network of question specifications provides guidance for sampling of the domain. Question specifications also allow several more questions to be generated for each specification than will actually be included on the final form or forms of the test. Generating additional questions is important because the statistical behavior of a single question, in relation to all other questions, defies prediction in advance. Questions generated from the same specification are treated as interchangeable thereby allowing two or more forms of a standardized test to be developed simultaneously.

The questions representing each question specification are not selected by a completely random process. This is because the each question specification must be represented in all final forms of the test. Random selection is an implicit assumption of the Domain Sampling model. However, due to the fact that the statistical performance of individual questions, in relation to all other questions, cannot be predicted in advance, the randomization requirement is at least partially satisfied by psychometric methods of question selection.

The assumptions of the domain sampling model regarding content sampling breaks down when applied to instructor-generated examinations for several reasons. The most glaring of these is that question specifications are rarely used to generate a large pool of questions from which to distill an examination. To overcome this weakness, we need to develop a different approach for determining inaccuracies related to content sampling for instructor-generated examinations. This is largely accomplished by controlling the amount of instruction occurring between two consecutive measurements using the concepts introduced in Chapter 1 concerning the definition of a sequence of instruction, abbreviated SOI. When using the SOI concept, content sampling error is controlled by including one or more questions for each piece of content contained in an SOI on the examination. In theory, this should reduce content sampling errors to zero. To insure adequate content sampling for a course or a collection of courses, these are divided into several relatively short SOI's. This has the effect of limiting the length of the examination required to intensively measure a student's understanding of the content of each SOI.

3.4 Mis-classification Errors

There are two types of mis-classification errors that are important for our purposes. The first is known as a Type I error and the second is known as a Type II error. A Type I error occurs when a student who actually possesses adequate understanding of the content of an SOI or of a domain fails to demonstrate this fact during assessment. The actual error occurs not when the student fails to perform satisfactorily on the assessment but instead, occurs when the student is not given credit for successfully possessing the knowledge. A Type II error occurs when a student who possesses an inadequate level of understanding of an SOI or of a domain produces a satisfactory score on the assessment and is given credit for possessing the knowledge.

A Type I error, which has a probability of occurrence equal to , is corrected by reassessing the student either in a similar way or in an alternative way that allows the student to demonstrate that they have the knowledge. A Type II error, which has a probability of occurrence of , is detected only when, subsequent to the assessment, the student demonstrates actions contrary to the instruction or shows incomplete understanding of the content of an SOI. Type II errors have serious implications for the organization because detection occurs in the post-instruction environment and often at considerable expense.

The level of understanding that is deemed satisfactory is estimated in one of two ways and which is selected depends upon whether a minimum allowable score exists in advance of the measurement process. For example, GED sub-tests have cutoff scores which are typically dictated by the state in which the test is taken. When an SOI is in support of a minimum allowable score on an external test, the proportion questions answered correctly required to demonstrate adequate knowledge of the content is not the minimum acceptable score on the external test. For our purposes, the minimum acceptable score on the examination for an SOI is substantially higher than the minimum allowable score on the external test. This is the mechanism we will use to reduce the Type II error rate to a low level. The details of selecting a score that differentiates acceptable from unacceptable performance on an examination in support of an external test score is discussed in detail in Chapter 4. For the purposes of this chapter, Type II error rates in the 1% to 2% range are desirable.

When an SOI is offered that is not in support of an external test score, a high proportion of questions must be answered correctly to receive credit for obtaining the knowledge. In certain circumstances, as is in certain types of safety training or training in critical protocols, where Type II errors are costly, the goal is to reduce the Type II error rate to 0. For these types of instruction, the goal is for all students to achieve perfect scores on the examination for all SOI's in a course or collection of courses.

The Type I error rate influences the number of students requiring reassessment or retraining. Large reassessment or retraining rates dramatically increase the cost of offering an SOI. In general, when the proportion of questions answered correctly is high, both the Type I and Type II error rates are minimized.

Regarding mis-classification errors on standardized tests, a student's actual knowledge of a domain is estimated by the obtained score. The obtained score is in-turn viewed as a proxy for the theoretical score the student would obtain if questions for each detail of the domain of interest could be included on a single test. This theoretical score is known as the student's 'true score' and deviations between the obtained score and the 'true score' are attributed to all known and unknown sources of variation that combine to obscure the student's ability to obtain a score equivalent to their 'true score'. This conception is known as the True Score Model.

When a student's obtained score varies widely from the student's 'true score', a mis-classification error can result. With respect to standardized tests, the Type I and II error rates are generally neither known nor knowable because the 'true score' is a theoretical quantity. Standardized tests instead rely upon test design techniques such as long test lengths, which provides more total information, and psychometric question selection, to minimize Type I and II errors. In a later section of this chapter, it will be shown that questions of intermediate difficulty maximize the ability of the test to discriminate between students having different 'true scores'. This in turn, minimizes mis-classification errors by allowing distinctions between students having different levels of knowledge of a domain to be consistently differentiated.

On an instructor generated examination for which a minimum allowable score exists, the test must be long enough to allow to be estimated and to minimize content sampling errors. If an SOI is long and detailed, an unacceptably long examination would be required in order to minimize content sampling errors. This implies that the optimal length of an SOI is influenced by the need to minimize content sampling errors, on the one hand, while having examinations that are long enough to resolve the level of detail necessary to detect Type II errors on the other.

3.5 Guessing Related Error

The final source of inaccuracy in Equation 3.2 is the influence of guessing behavior. The effects of guessing behavior on obtained scores has been studied extensively so certain informed decisions are possible regarding the control of this source of inaccuracy. Inaccuracies related to guessing are generally the largest source of inaccuracy for instructor generated examinations once errors related to content sampling and mis-classification have been minimized. The influence of guessing behavior on standardized tests can be large but is believed to be small in comparison to content sampling errors.

Regarding standardized tests, even though content sampling errors are the largest source of error, two strategies are used in combination to reduce the influence of guessing on obtained scores. The first of these is an explicit correction for guessing which consists of raising the target proportion of correct answers for the test. In general, standardized tests are maximally discriminating between students of differing abilities at the target proportion of questions answered correctly used during psychometric pruning of the initial questions. The target proportion of questions answered correctly is raised enough to account for the expected number of correct guesses the average student will make on the test. More will be said about this later in this chapter. The second strategy is developing tests of sufficient length to allow the random nature of guessing to cancel out the effect of guessing on obtained scores. This implies that the length of a standardized test must not only be long enough to adequately sample the domain of interest, but also long enough to neutralize the influence of guessing on obtained scores.

Instructor-generated examinations are typically not long enough to neutralize the effects of guessing on obtained scores. The influence of guessing behavior on examinations is best controlled by using question formats that reduce the chances of guessing the correct answer. The rationale for using question format as a strategy for controlling the influence of guessing on instructor generated examinations is the mathematical proof shown in Equation 3.3. In this equation, each question has a probability of being guessed correctly of less than or equal to P_g. The contribution of each question to the overall %inaccuracy equals the P_g value for the question divided by the total number of questions on the examination. An examination comprised of these questions will have an overall level of inaccuracy related to guessing of less than or equal to the sum of the contributions from each question.

Due to the goal of having total inaccuracy contained in each measurement of 3% or less, if content sampling is assumed to be near 0% and mis-classification errors are 1%, the inaccuracy introduced by guessing behavior must be less than 2%. The implication is that in order for inaccuracies due to guessing to be 2% or less, the number of viable response options for each question must be large in order to minimize P_g. Types of question formats that can be written to provide a large number of viable response options are free response, matching and fill in the blank to name a few. For questions using a matching format, instructions indicating that response options may be used more than once helps maximize the number of viable response options.

Equation 3.3: Inaccuracy Related to Guessing

$$\mathit{Guessing}\ \%\ \mathit{Inaccuracy} = \frac{\sum \%\ \mathit{per\ Question}}{\#\ \mathit{Questions}}$$

3.6 Guessing Related Error on Standardized Tests

On standardized tests, the use of multiple choice questions having a small number of viable response options is rationalized by the fact that content sampling errors are believed to be large in relation to the influence of guessing. Multiple choice questions typically have either four or five response options and so have a minimum probability of being guessed correctly equals $P_g = 0.25$ or 0.20, respectively. P_g is known to be underestimated when considering only the number of response options because of the way in which multiple choice questions are answered by students.

A functioning multiple choice question contains three types of response options where each type of response option has a distinctly different probability of being selected when guessing. The first type of response option is the correct answer. The second type is a response option, known as a distractor, that is often partially correct. The third type of response option is either unrelated or weakly related to the correct response and is known as a foil. A typical multiple choice question contains a correct response, a single distractor and two or three foils. In practice, students' have learned to eliminate most or all of the foils before guessing. This is known as using partial knowledge. This results in the actual probability of guessing correctly as high as $P_g = 0.5$ if all foils can be eliminated before guessing between the distractor and the correct answer. A probability of guessing correctly of $P_g = 0.33$ occurs when all but one foil and the distractor can be eliminated before guessing.

The theoretical point of maximal discrimination for tests developed using psychometric methods, when no error related to guessing is considered, is obtained when the average probability of answering a question correctly equals $p = 0.5$. At $p = 0.5$, the variation of the binomial distribution is maximized because the variance equals $p \times q = p \times (1 - p) = 0.5 \times 0.5 = 0.25$. No other combination of p and q will yield a larger variance.

Maximizing the variation captured by each question results in a maximally discriminating test. However, in the face of guessing behavior, an average probability of answering a question correctly of p = 0.5, when an increment to account for guessing is added, results in a point of maximal discrimination of p > 0.5. To compensate, the average probability of answering a question correctly is increased to account for guessing behavior using Equation 3.4. In light of Equation 3.4, the question becomes whether to include an estimate of the average partial knowledge used before guessing in the explicit correction for guessing to the test p.

For example, a maximally discriminating test when strong partial knowledge is operational results in having an average probability of answering a question correctly of p = 0.5 + 0.5 x (1 - 0.5) = 0.75 regardless of the number of response options. This is because all but two responses are always eliminated from consideration before guessing. Alternatively, if the assumption is that no partial knowledge is operational and the number of response options for all question on the test equals five, a maximally discriminating test occurs when the average probability of answering a question correctly equals p = 0.50 + 0.20 x (1 - 0.5) = 0.6. Other target values of p between these two extremes of p = 0.75 and p = 0.60 are also possible and depend solely on assumptions about partial knowledge. The necessity of including an explicit correction for guessing into the standardized test development process has had an unintended benefit for the K-12 standardized achievement test industry. The correction for guessing allows tests of moderate difficulty, of say p = 0.7 or even p = 0.75, to be maximally discriminating. Tests having difficulty levels approaching p = 0.75 are better tolerated by test takers than tests having an average difficulty of p = 0.5.

Equation 3.4: p Correction to Standardized Test

$$Corrected\, p = p + P_g \times (\mathbf{1} - p)$$

The popularity of standardized tests comprised of four or five option multiple choice questions has resulted in this question format becoming extremely popular for instructor-generated examinations. This is problematic for several reasons. First, due to the fact that P_g is large for this question format, inaccuracies related to guessing become the overwhelming source of inaccuracy. Second, due to a desire to minimize content sampling errors, examinations are rarely long enough for the random nature of guessing to neutralize the influence of guessing on obtained scores. Finally, average proportions of questions answered correctly in the range of p = 0.6 to 0.75 on an instructor generated examination, as opposed to a standardized test, result in large Type II error rates.

It is possible to reduce the influence of guessing upon obtained scores for both standardized tests and instructor generated examinations by subtracting the expected number of correct guesses from the obtained score. This is accomplished by estimating the number of correct guesses from the number of questions missed then subtracting this quantity from the obtained score. It should be noted that an after-the-fact correction for guessing, as a means of reducing the influence of guessing behavior on an obtained score, does not provide the theoretical assurance afforded by Equation 3.3.

Specifically, Equation 3.5 is used for estimating the number of correct guesses from the number of questions missed. The estimated number of correct guesses is then added to the number of questions a student misses (W). The original correction for guessing formula shown in Equation 3.5 tends to underestimate the impact of guessing when students use partial knowledge to eliminate responses before guessing. In the original formulation, the chances of guessing correctly equals 1 / A, where A is the number of response options.

Equation 3.5: Original Scoring Rule for Guessing

$$Guesses = \quad G = W + \frac{W}{A-1}$$

In order to introduce the concept of partial knowledge into Equation 3.5, the chance of guessing a question correctly becomes A_E where A_E = A - (the number of response options eliminated before guessing) as shown in Equation 3.6.

Equation 3.6: Modified Scoring Rule for Guessing

$$G\ with\ Partial\ Knowlwdge = W + \frac{W}{A_E - 1}$$

For example, if strong partial knowledge is operational, students can always eliminate all but two response options before guessing. Under the assumption of perfect partial knowledge, Equation 3.6 is implemented as a scoring rule stating that one correct answer will be subtracted from the total number of correct answers for each question missed as shown in Equation 3.7.

Equation 3.7: Scoring Rule for Strong Partial Knowledge

$$G\ with\ Strong\ Partial\ Knowledge = W + \frac{W}{2-1}$$

3.7 Guessing Related Error on Instructor Generated Examinations

In order to explain the usefulness of Equation 3.3, what is needed is a model that more accurately reflects the effects of guessing behavior on obtained scores for examinations. In the model contained in Figure 3.1, due to the random nature of guessing after all partial knowledge has been exhausted, guessing behavior conforms to a binomial distribution defined by the number of opportunities to guess and the probability of guessing correctly during a certain number of opportunities to guess. The binomial distribution is formed by employing the counting rules of probability theory to exhaustively determine the probability of all possible ways a certain number of correct guesses can occur taken a predefined number of questions at a time. The expected number of correct guesses occurs at P_g times the number of opportunities to guess. Less probable patterns of correct guesses also occur. For example, a rare pattern for 20 opportunities to guess occurs when 17 guesses are successful. In counting rule language, this is equivalent to the expression 17 correct guesses taken 20 questions at a time. If strong partial knowledge is operation, P_g = 0.5 and the probability of 17 correct guesses in 20 opportunities is quite small equaling 0.0011.

The guessing model for P_g = 0.5 is illustrated in the top portion of Figure 3.1. In this figure, the obtained score is comprised of two components, the Lower Bound score and the number of correct guesses. The Lower Bound score equals the obtained score, expressed as a proportion of the total number of questions, once the expected influence of guessing has been completely removed. Equation 3.8 is used to calculate the Lower Bound score. This conception is critical because the Lower Bound score is exactly what we want to know concerning a students knowledge of the content of an SOI. In practice, content sampling error and the Type II error rate are imbedded in the Lower Bound score and cannot be fractioned out. However, if content sampling errors are ~ 0% and ~ 0.01, the proportion of the obtained score that can be attributable to guessing behavior must be less than ~ 0.02.

Figure 3.1: The Obtained Score = Lower Bound Score + Number of Correct Guesses.

$$\overbrace{\left(\textit{Lower Bound Score}\right)\left(\textit{Correct Guesses}\right)}^{\textit{Obtained Score}}\left(\textit{Incorrect Guesses}\right)$$

Equation 3.8: Lower Bound Score

$$\textit{Lower Bound Score} = \frac{\left(\textit{Score} - P_g\right)}{\left(1 - P_g\right)}$$

For example, if an examination contains 20 matching questions where there are always 10 viable response options, P_g = 0.1. The Lower Bound score for a student having an obtained score of 16 of 20 questions equals (0.8 - 0.1) / (1 - 0.1) = 0.7 / 0.9 = 0.78. Notice how close the Lower Bound score of 0.78 is to the obtained score of 0.80. Rounding down to the next possible score lower than 0.78 yields a score of 15 of 20 correct or p = 0.75. After rounding down, there are 20 - 15 = 5 opportunities to guess.

In Equation 3.9, the parameters used to select the correct binomial distribution for a particular number of opportunities to guess are the number of trials, which equals the number of opportunities to guess (n), and a probability of a success equal to $P_{g.}$ The number of successes equals the number of correct guesses (m).

Continuing the discussion, the probabilities of guessing m = 0, 1, 2, 3, 4 or 5 questions correctly on an n = 20 question test are obtained by referring to a table of binomial probabilities. The probabilities for n = 20, m = 0, 1, 2, 3, 4 and 5 and P_g = 0.1 are: 0 (0.59049), 1 (0.32805), 2 (0.07290), 3 (0.00810), 4 (0.00045) and 5 (0.0001). The minimum increments of guessing behavior on a 20 question examination = 1 /20. The minimum increment of guessing for a particular number of correct guesses, m, equals (1 / 20) x m.

For example, for 0, 1, 2, 3, 4 and 5 opportunities to guess the minimum increment of guessing equals (1/ n) x m = (1 / 20) x m = 0.0, 0.05, 0.10, 0.15, 0.20 and 0.25. So, P(Inaccuracy) = 0.00(0.59049) + 0.05(0.32805) + 0.10(0.07290) + 0.15(0.00810) + 0.20(0.00045) + 0.25(0.0001) = 0.00000 + 0.01640 + 0.00729 + 0.00122 + 0.0009 + 0.00000 = 0.025. 100 x 0.025 = 2.5% inaccuracy related to guessing. Remember that we rounded down to obtain a whole number of correct guesses so, the %Inaccuracy is < 2.5%.

An interesting point is that for these parameters, a hypothetical student has a 59.05% chance of guessing 0 questions correctly and a combined 59.05% + 32.80% = 91.85% chance of guessing no more than 1 question correctly. Stated another way, the student has a ~ 92% chance of having only 0.0000 + 0.0164 = 0.0164 or 1.64% inaccuracy related to guessing.

Equation 3.9: Inaccuracy Related to Guessing

$$P(Inaccuracy) = \sum_{k}^{i=1} \left(Pw_i \times \left(\frac{1}{n} \times m_i \right) \right) \ where,$$

$$\begin{aligned} Pw_i &= binomial\ probability\ of\ a\ specific \\ &\quad\ number\ of\ correct\ guesses \\ m_i &= number\ of\ correct\ guesses\ at\ i \\ n &= number\ of\ total\ questions \end{aligned}$$

In summary, it is desirable for each student to have an obtained score very close to the Lower Bound Score. The key to this happening is to have the correct combination of three variables. The first is to minimize the number of opportunities to guess. This is accomplished by having each SOI bring each student to a high level of understanding of the content of the SOI. This should result in high obtained scores which minimize the opportunities to guess. As mentioned earlier, high obtained scores also have the effect of minimizing Type I and II errors. Second, as the number of questions on the examination increases, the minimum increment of guessing error decreases. This implies that extremely short SOIs, for which an examination of ~10 questions eliminates content sampling errors, are too short to bring the guessing error to <2%. Third, the number of viable response options for each question should be large because this minimizes P_g. The next two examples should illuminate the preceding conclusions.

Example 3.1

In this first example, a student answers 16 of 20 five-option multiple choice questions correctly where strong partial knowledge is operational. So, $P_g = 0.5$, p = 16 / 20 = 0.8 resulting in a Lower Bound score = (0.8 - 0.5) / (1 - 0.5) = 0.6 or 0.6 x 20 = 12 questions correct. Notice that the difference between the Lower Bound score and the obtained score is quite large equaling 4 questions.

The increment of error related to correctly guessing a single question on a 20 question examination equals 1 / 20 = 0.05. Using the minimum increment of guessing and the binomial probability of a certain number of correct guesses results in the solution contained in bottom row of Table 3.1. The sum equals P(Inaccuracy) = 0.000 + 0.002 + 0.011 + 0.033 + 0.055 + 0.055 + 0.033 + 0.0110 + 0.002 = 0.20. For example, the value 0.011 for two correct guesses is calculated by multiplying 0.10 x 0.109 = 0.109 ~ 0.11. In this example, the error attributable to guessing sums to 0.20 or 20%. Having 20% inaccuracy introduced into an obtained score as a result of guessing behavior is a direct result of the question format in conjunction with the assumption that strong partial knowledge is operational.

Table 3.1: Inaccuracy Due to Guessing for P_g = 0.5

Guesses	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Comb.	1	8	28	56	70	56	28	8	1
Prob.	0.004	0.031	0.109	0.219	0.273	0.219	0.109	0.031	0.004
Min.	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40
Pr. Error	0.000	0.000	0.011	0.033	0.055	0.055	0.033	0.011	0.002

A further use of Table 3.1 is to calculate the probability of a specific range of obtained scores such as a score of 16 +/- 1 questions correct. An obtained score of 16 is possible if 4 correct guesses are added to the Lower Bound score of 12. In order to accomplish this, the combined probabilities for 3, 4, and 5 correct guesses are added. The sum of these probabilities equals 0.219 + 0.273 + 0.219 = 0.711 or a 71.1% chance. A further use of Table 3.1 is to calculate the probability of obtaining a score above the Lower Bound score. This equals the probability of an obtained score of 13 questions or greater. Due to the large probability of guessing correctly of P_g = 0.5, this probability equals 1 - 0.004 = 0.996 or a 99.6% chance.

The measurement system described in Example 3.1 clearly introduces too much inaccuracy into each measurement. In this next example, all of the information is the same except the number of viable response options, after partial knowledge is considered, equals 15 response options. This reduces the probability of guessing the correct answer to P_g = 1 / 15 = 0.067. In Example 3.2, we will see that this single change has a dramatic effect on the %Inaccuracy introduced as a result of guessing behavior.

Example 3.2

In this second example, the students' obtained score again equals 16 of 20 questions so the obtained score equals p = 16 / 20 = 0.80 and the minimum increment of guessing behavior equals 1 / 20 = 0.05. The examination is comprised of matching format questions in which there are no less than 15 viable response options for each question and so P_g = 1 / 15 = 0.067. The Lower Bound score equals (0.80 - 0.067) / (1 - 0.067) = 0.786. Using this information, the Lower Bound score equals 0.786 x 20 = 15.72 questions. Rounding down to insure conservatism, the Lower Bound score equals 15 questions. This results in 20 - 15 = 5 opportunities to guess. In Table 3.2, the second row contains the binomial probabilities corresponding to each number of correct guesses out of 5 opportunities to guess. As you can see, the student has a probability of 0.708 of producing an obtained score equal to the Lower Bound score. The percent inaccuracy related to guessing behavior for a 20 question examination comprised of questions having 15 viable response options and 5 opportunities to guess equals 0.00 + 0.0127 + 0.0036 + 0.0004 + 0.00 + 0.00 = 0.0167 ~ 1.7%. If inaccuracies related to content sampling are near zero and the Type II error equal 0.01, the total inaccuracy to known sources equals 0.01 + 0.017 = 0.027 or 2.7%. Considering that we rounded down to the nearest whole number of guesses means %Inaccuracy < 2.7% . This is within the 3% or less expectations for the %Inaccuracy.

Table 3.2: Inaccuracy Due to Guessing for $P_g = 0.067$

Guesses	**0**	**1**	**2**	**3**	**4**	**5**
Prob.	0.708	0.253	0.036	0.003	0.000	0.000
Min.	0.00	0.05	0.10	0.15	0.20	0.25
Pr. Error	0.0000	0.0127	0.0036	0.0004	0.0000	0.0000

3.8 Item Analysis

As with the development of an observational instrument, the development process for both instructor generated examinations and standardized tests begin by developing a large number of questions called an item pool. Generating an item pool for an instructor generated examination is often a less detailed process than that used for development of a standardized test. This owes mainly to the complexity of uniformly sampling a domain.

Regardless of whether an examination or a test is being developed, a few more questions are generated than will be required for the initial form or forms of the assessment. When developing a standardized test, many questions are generated from each question specification. Each question specification must be represented in all forms of the test after psychometric pruning of the item pool. For instructor-generated examinations enough questions are generated to produce two or three forms of an examination. It is important that all questions are perceived as fair by the majority of test or examination takers. To this end, questions should be accurate, have appropriate readability and be unbiased.

After the item pool is complete, the development process continues by administering the item pool to ten or more students having the mix of abilities present in the target population. In the development of an examination, two or three questions for each piece of content in an SOI are administered to actual students.

A goal of instruction employing SOI concepts for reducing content sampling errors is to have any fair question answered correctly by all students. A goal of psychometric methods is to produce a test that has a target proportion of questions answered correctly where the test taker answers as many questions as their knowledge of the domain allows. Typically the target proportion of correct answers contains an explicit correction for guessing. Unfair questions, also known as flawed items, should be removed from the final form or forms of either type of assessment. One way to identify flawed items is by performing a set of procedures collectively known as Item Analysis.

Items that are answered correctly by only a few students often contain a flaw in the statement portion of the item. A second indicator of a flawed item occurs when the item scores are negatively correlated with the total scores. A negative correlation indicates that the item is answered correctly by a large number of individuals who score poorly on an examination or test as a whole. There is one item analysis pattern that is specific to standardized tests. This occurs when an item is either extremely difficult, in the absence of a mis-leading statement, or extremely easy. These questions contribute little information, in a psychometric sense, and should be reworked or eliminated from the initial forms of a test. There are three additional undesirable patterns: guessing, mis-keying and ambiguity. These flaws are detected by tallying the number of times each response option is selected for a particular item and looking for patterns of responses indicating the undesirable patterns.

Guessing is indicated when most response options, in a multiple choice question format, are selected at approximately the same frequency. In cases where one or more of the foils can be eliminated through the use of partial knowledge, an approximately equal frequency of selection for the distractor and correct response may also indicate guessing. For question formats designed to minimize P_g, this pattern is indicated when several response options are selected at the same frequency. In mis-keying, a single incorrect response is selected at a higher frequency than the correct response. An item which is ambiguous, either in the statement portion of the item or in the response options, will have most responses approximately equally divided between two response options. This would indicate that both responses are either equally correct or that both responses are acting as distractors in that neither response is completely correct. Examples of these response patterns are shown in Table 3.3. Example 3.3 illustrates these item analysis concepts.

Table 3.3: Flawed Response Patterns (B is the correct response)

Responses	**A**	**B**	**C**	**D**	**E**
Guessing	7	8	9	8	3
Mis-keying	1	9	21	2	3
Ambiguous	1	15	15	2	3

Example 3.3

The following example evaluates the item pool for a sub-test which should ultimately contain six or seven items after psychometric pruning. Table 3.4 contains the initial thirteen questions, the correct answers, the reason for rejecting an item, the response frequencies and finally the item difficulty. The item pool was exposed to a group of 20 students having mixed levels of ability.

Of the 13 items, items 5, 7, 11 and 13 appear to be extremely difficult. Upon further evaluation, these four items have response frequency patterns indicative of flaws in the item's performance. Item 5 has three response options selected at about the same frequency. This pattern indicates that the statement portion of the item is ambiguous and fails to distinguish between response options B, C and E. In items 7 and 13, all response options are selected at approximately the same frequency. This indicates that the statement portion of the item fails to identify the correct response based upon their knowledge so test takers are simply guessing. Item 11 is extremely difficult but only one response is selected at a high frequency. This patterns indicates a mis-keying error and once this clerical error has been corrected, the item is included in the initial sub-test.

Once the three flawed items have been removed from the item pool and the mis-keyed item has been corrected, the obtained scores for each question for each of the 20 students exposed to the sub-scale under development are contained in Table 3.5. A score of 1 means the question was answered correctly and a score of 0 means the question was answered incorrectly. The table is sorted in descending order using the rightmost column which contains the total score for each student. The bottom two rows contain the proportion of correct answers for each question followed by the correlation coefficients comparing each question to the total scores. The effect of the question being correlated has been removed from the total scores prior to calculating the correlation in the same fashion used in the development of an observational instrument.

Table 3.4: Item Analysis

Item	Answer	Error	A	B	C	D	E	p
1	C		1	4	13	1	1	0.65
2	C		0	2	14	4	0	0.70
3	D		0	1	0	17	2	0.85
4	E		0	2	4	1	13	0.65
5	B	Ambiguous	0	7	7	1	5	0.35
6	B		4	10	4	2	0	0.50
7	C	Guessing	4	5	5	3	3	0.25
8	A		15	4	1	0	0	0.75
9	A		17	0	2	1	0	0.85
10	A		14	0	6	0	0	0.70
11	E	Mis-keyed	0	0	0	16	4	0.20
12	E		1	0	0	1	18	0.90
13	D	Guessing	5	4	4	3	4	0.15

The idea is to have all items significantly positively correlated with the total scores. Using Equation 3.10, the size of the correlation coefficient required for significance for 20 - 2 = 18 degrees of freedom and = 0.05 equals 0.444 as shown in Equation 3.11. In the initial sub-test, items 3, 6, 8, 11 and 12 are not significantly correlated with total scores. The correlations for items 3 and 12 are especially low. If an examination is being developed, items 3 and 12 are removed and items 6, 8 and 11 are re-evaluated once more data are available. For a standardized test, many or all of these items may be eliminated during psychometric pruning. If this leaves an insufficient number of questions or if an item specification will be eliminated from the test or sub-test, more items are generated and the study is repeated.

Equation 3.10: *r* Required for Significance

$$r = \frac{t_{n-2,\alpha}}{\sqrt{(n-2)+t^2_{n-2,\alpha}}}$$

Equation 3.11: *r* for =0.05

$$r = \frac{2.102}{\sqrt{(20-2)+2.102^2}} = 0.444$$

Table 3.5: Correlations between Question Scores and Total Scores

Ques.	**1**	**2**	**3**	**4**	**6**	**8**	**9**	**10**	**11**	**12**	**Total**
Pat	1	1	1	1	1	1	1	1	1	1	10
Rose	1	1	1	1	1	1	1	1	1	1	10
Jill	1	1	1	1	1	1	1	1	1	1	10
Jim	1	1	1	1	1	1	1	1	1	1	10
Linda	1	1	1	1	1	1	1	1	1	1	10
Joe	1	1	1	1	1	1	1	1	1	1	9
Sam	1	1	1	1	1	1	1	1	0	0	9
Don	1	1	1	1	0	1	1	1	1	1	9
Nick	1	1	1	1	1	0	1	1	1	1	9
Khoi	1	1	1	0	1	1	1	1	1	1	9
Marty	1	1	1	0	0	1	1	1	1	1	8
Hai	1	1	1	1	0	0	1	1	1	1	8
Ray	1	1	0	1	0	1	1	1	1	1	8
Lee	0	1	1	0	0	1	1	1	1	1	7
Sam2	0	0	0	1	0	1	1	0	1	1	5
Ty	0	0	0	1	0	1	1	0	1	0	4
Patty	0	0	1	0	0	0	1	0	0	1	3
Julie	0	0	1	0	0	1	0	0	0	1	3
Pam	0	0	1	0	0	0	0	0	1	1	3
Joey	0	0	1	0	1	0	0	0	0	1	3
P	*0.65*	*0.70*	*0.85*	*0.65*	*0.50*	*0.75*	*0.85*	*0.70*	*0.80*	*0.90*	
r	0.89	0.92	0.13	0.47	0.43	0.32	0.60	0.92	0.41	-0.01	

3.9 Objectivity of Tests and Examinations

Now that the issues surrounding the accuracy of obtained scores have been presented, our discussion turns to topics related to the objectivity of obtained scores produced by measurement systems based around examinations or tests. As shown in Equation 2.3, objectivity is comprised of two non-overlapping components known as Repeatability and Reproducibility. As was the case with accuracy, determining the objectivity of measurement systems based around either an examination or a test is hampered by the irreversible alteration of students resulting from exposure to the assessment. Repeatability is not impacted to the same degree by the irreversible alteration of the test taker as is reproducibility. This is because there are existing methods for estimating the repeatability component of objectivity for measurement systems based around these two types of assessment. Regarding reproducibility, irreversible alteration can make the statistical independence of reproduced measurements suspect.

Objectivity estimation procedures for examinations and tests both use an active in-situation objectivity study similar to that used for an observational measurement system. This is because the assumptions required for use of the bivariate normal distribution can be met by the procedures for developing two or more statistically identical tests or examinations. The two or more statistically identical forms of the same examination or test, known as parallel forms, are developed simultaneously. This allows the correlation between the two forms to be calculated and used to estimate the repeatability component of objectivity.

As discussed earlier, there are two broad categories of assessments, instructor-generated examinations and standardized tests. At this point, the pathways for developing these two types of assessments diverge. We will continue with procedures for developing instructor-generated examinations before moving to a discussion of the psychometric methods used in the development of standardized tests.

3.10 Objectivity of Instructor Generated Examination

As already discussed, we know that measurements that are both accurate and objective contain the most extractable information. The goal of this discussion is to illustrate a uniform and flexible method of developing parallel forms of an instructor-generated examination which have low inaccuracy and high objectivity. In all instructor generated examinations, SOI concepts are used for controlling content sampling errors and mis-classification errors are minimized by requiring high obtained scores. Question formats are also used to reduce the %Inaccuracy attributable to guessing to less than 2%.

The objectivity of parallel forms is calculated using Equation 2.4 for calculating the correlation coefficient and Equation 2.5 for calculating the discrimination index D. The interpretation of D is consistent with Figure 2.2 where $D \geq 10.0$ is desirable. Due to the risk of introducing uncontrolled variables into the examination development process, two forms of the examination are administered in close proximity either by administering one form followed directly by the second form or by mixing the two forms into one large examination to be separated later.

The examination development process for parallel examinations begins by generating pairs or triplets of questions. If two forms of the examination are desired, the member of each triplet having the lowest correlation to the other two members of the triple is eliminated. One member of the resulting pairs of question is randomly assigned to one form of the examination and the other pair member to the other form of the examination. Random assignment of each pair member helps to avoid question selection bias and should result in two approximately equivalent forms of the examination. Short of generating question specifications, every attempt should be made to write pairs or triplets of questions that are interchangeable in a manner analogous to the process used to generate question when question specifications are available.

After both forms of the examination are complete, both forms are administered to a representative group of students and the scores for the two forms are correlated. The smallest correlation coefficient that is considered adequate, between the two forms of the examination, equals r = 0.981 which corresponds to a D value of 10.2. Example 3.4 illustrates the concepts necessary for developing an instructor generated examination.

Example 3.4

In the following example, 43 students were administered an examination comprised of triplets of questions, believed to be interchangeable within each triplet, covering each piece of content in an SOI. The goal is to develop two final forms of the examination containing 24 total questions where each examination is comprised of four sub-tests. The correlations among all possible pairs for the first three triplets are shown in Table 3.6. The correlation coefficient between two dichotomus samples is known as Phi and abbreviated . Using Equation 3.12, the critical value required for significance at = 0.05, N = 86 equals $1.96 / (86 - 1)^{0.5} = 1.96 \times 0.1085 = 0.213$. In the table, all of these items are significantly correlated at = 0.05. The pairs of items having the highest level of correlation for each of the first three triplets are 1a and 1c, 2a and 2b and 3a and 3b. These pairs are selected for the final forms of the examination.

Equation 3.12: Significance of at = 0.05 and = 0.01

$$\phi_{.05} = \frac{1.96}{\sqrt{N-1}} \text{ or } \phi_{.01} = \frac{2.58}{\sqrt{N-1}}$$

Table 3.6: Inter-correlation Matrix for Item Triplets 1, 2 and 3

	1b	1c	2b	2c	3b	3c
1a	0.64	0.88				
1b		0.46				
1c						
2a			0.85	0.71		
2b				0.45		
2c						
3a					1.00	0.85
3b						0.85

After all of the triplets have been reduced to two questions, the obtained scores for Form A and Form B are shown in Table 3.7. Notice that the average score for Form A, which equals 19.72, is in close agreement with that of Form B which equals 19.77. The variance for each form are similar equaling 4.43 and 4.50 for Form A and B, respectively. However, even though the two form are nearly identical, the correlation between the two forms equals r = 0.886 which is well below the r = 0.981 required for acceptable objectivity. The low correlation is a direct result of the discrete nature of the question scores.

Table 3.7: Correlation Between Obtained Scores for Forms A and B

Stu.	A	B	**Stu.**	A	B	**Stu.**	A	B	**Stu.**	A	B
1	16	17	**12**	18	18	**23**	21	18	**34**	22	23
2	14	15	**13**	17	17	**24**	20	18	**35**	22	23
3	18	19	**14**	20	20	**25**	18	19	**36**	21	22
4	17	17	**15**	20	21	**26**	20	20	**37**	21	21
5	16	16	**16**	18	19	**27**	21	22	**38**	22	22
6	17	18	**17**	18	18	**28**	19	20	**39**	22	22
7	18	18	**18**	18	18	**29**	21	19	**40**	23	24
8	19	19	**19**	20	18	**30**	22	21	**41**	22	23
9	20	21	**20**	21	19	**31**	22	21	**42**	24	24
10	20	21	**21**	21	20	**32**	20	20	**43**	22	22
11	18	18	**22**	20	20	**33**	19	19			
									Avg.	19.72	19.77
									Var.	4.434	4.504
									r	0.886	

As a result of the inadequate correlation between Forms A and B, each test is divided into the four sub-tests as shown in Table 3.8. Form A is comprised of sub-tests A1 to A4 and Form B is comprised of sub-tests B1 to B4. The averages for each sub-test are shown in the last row of the table. The correlation between the averages for the sub-test scores for Forms A and B equals r = 0.993 as shown in Table 3.9.

Table 3.8: Sub-test Scores

Stu.	**A1**	**A2**	**A3**	**A4**	**B1**	**B2**	**B3**	**B4**
1	4	5	4	5	4	5	4	6
2	4	4	4	5	4	4	4	5
3	4	4	4	4	4	5	4	3
4	4	4	4	5	4	5	4	5
5	4	5	4	5	4	4	4	6
	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.
39	5	6	5	6	5	6	5	6
40	5	5	5	8	6	6	6	6
41	5	6	5	6	5	5	5	8
42	6	6	6	6	6	7	6	5
43	5	6	5	6	5	6	5	6
	4.60	5.07	4.60	5.40	4.53	5.21	4.53	5.49

Table 3.9: Sub-test Correlations

Form A	**Form B**
4.60	4.53
5.07	5.21
4.60	4.53
5.40	5.49

r = 0.993

The D value for r = 0.993 equals 16.9 as shown in Equation 3.12. This D value produces a rating between acceptable and good according to Table 2.2. However, this means that the resolution of these two examinations is limited to the sub-test level rather than the individual question level. This will often be the case for instructor generated examinations.

Equation 3.12: Discrimination Index for r = 0.993

$$D = \sqrt{\frac{1 + 0.993}{1 - 0.993}} = 16.9$$

3.11 Objectivity of Standardized Tests

Standardized test development is a lengthy and often expensive process and should only be undertaken after careful consideration. The application of psychometric methods does however offer a systematic approach for selecting the subset of questions from an item pool most likely to produce the same obtained score for a student, on two separate occasions without the first exposure to the test influencing the second exposure. Of course irreversible alteration of students by exposure to the test on the first occasion makes this impossible. There are, however, methods for estimating the quality of the test being developed requiring only a single exposure to the test. Even though a full treatment of the topic of psychometric test development is beyond the scope of this text, some understanding of how a standardized test is developed is vital for organizations employing standardized tests to assess gains in student performance.

Standardized tests are used in education and training settings to assess the knowledge a student possesses in a topic area so broad that writing enough questions to thoroughly examine each piece of information is impractical. There are two basic categories of standardized tests used for this purpose, achievement tests and norm referenced tests. A norm-referenced test attempts to determine a student's knowledge with respect to some population norm such as grade equivalents. There are many excellent commercially available norm-reference tests designed to measure knowledge in many areas of concern. It should be noted that these types of tests are some of the most useful and successful applications of psychometric theory. Due to the availability of excellent norm-referenced tests, our focus will be upon developing standardized achievement tests.

Many psychological constructs, such as intelligence, are of interest because of the variability with which these are expressed in the population at large. Achievement on a standardized test is not a direct measure of intelligence, but experience has shown the scores for students given an intelligence test and an achievement test are highly correlated. Once a domain has been identified, the portions of interest are described in a domain specification from which question specifications are generated. In theory, all questions generated from the same question specification are statistically interchangeable. Each question is then administered to a large and varied collection of students in order to provide data for determining the statistical behavior of each question in relation to all other questions. Typically, all questions on the initial forms of a standardized test are administered to ten or more students. Not all of the questions generated will perform well enough to be included in the final form or forms of the test. As already mentioned, it is impossible at the outset to predict which questions will survive the psychometric pruning process. This means that in order to produce a standardized test containing 50 questions, generating an item pool of around 150 questions is common.

Due to the relatively small number of question selected from the domain of interest, errors attributed to inadequate content sampling are believed to be by far the largest sources of inaccuracy. Since the amount of inaccuracy attributed to guessing is believed to be small relative to content sampling errors and the fact that standardized tests typically contain enough questions to allow the random nature of guessing to cancel out the effect of guessing on obtained scores, a special type of question format was developed for use in standardized tests namely, the multiple choice question. The multiple-choice question format is only truly appropriate in situations where content sampling errors are believed to be the overwhelming source of error and the number of questions on the assessment is relatively large.

Multiple choice questions typically have either four or five response options and as mentioned earlier, the response options for a well written multiple choice question consist of the correct response, two or three foils and a specialized foil called a distractor. Foils, other than the distractor, are obviously incorrect to students with some level of partial knowledge. These foils are eliminated from consideration by test savvy students. The distractor is nearly correct, in one aspect or another, and is designed to mislead students using partial knowledge before guessing. From a sufficiently large sample of students, the pattern of responses to the options on a particular question gives an indication as to whether the question is functioning correctly. This is verified using Item Analysis techniques.

The success of the psychometric test development process is summarized by one or more correlation measures collectively known as psychometric reliability coefficients. Psychometric reliability coefficients range from 0 to 1 where a psychometric reliability coefficient near 1.0 indicates that the test being constructed is likely to be objective. Most psychometric coefficients are derivations of the Pearson Product Moment correlation coefficient shown in Equation 2.5. Simplifying assumptions are made to the product moment equation in order to estimate psychometric reliability from a single administration of a test. One of the most straight forward of these methods is the Split-half method. In the Split-half method, an examinations is divided into half tests A and B. The half tests are administered to a representative group of students on a single occasion. Then the correlation coefficient is calculated comparing the two halves. This Split-half psychometric reliability coefficient is then corrected using a formula known as the Spearman-Brown correction, shown in Equation 3.13, in order to estimate the psychometric reliability of a single test comprised of both halves. In Equation 3.13, r_{AB} equals the Split-half correlation and r_{tt} equals the predicted psychometric reliability. For example, if the Split-half correlation equals 0.60, the corrected Split-half reliability equals r_{tt} = (2 x 0.6) / (1 + 0.6) = 1.2 / 1.6 = 0.75.

Equation 3.13: The Spearman-Brown Correction

$$Predicted\ r_{tt} = \frac{2r_{AB}}{1+r_{AB}}$$

The major complaint concerning the Split-half method for estimating psychometric reliability is that several methods have been proposed for splitting the test, all of which result in a different Split-half psychometric reliability. Some of the methods proposed for splitting a singe test into two halves are listed below. If this is the method of determining psychometric reliability selected by an organization, one way of splitting a test is selected and applied to all tests developed within the organization.

1 - Form one half from odd number questions and the other from even number questions

2 - Rank order the questions in terms of difficulty and assign odd ranked questions to one half and even ranked questions to the other.

3 - Randomly assign the questions to each half

4 - Generate pairs of questions from the same question specification and assign one question from each pair to each half.

Due to the issue of splitting a test in to equivalent halves, a simplified version of the Pearson Product Moment equation has been developed. This derivation assumes that the variance of samples x and y are equal. The resulting equation provides an estimate of the psychometric reliability for an intact test administered on a single occasion. The version of this equation for use with dichotomous data is known as Kuder-Richardson 20 and abbreviated KR_{20}. Specifically, KR_{20} is a version of Coefficient Alpha that accommodates dichotomus data. The formula for the KR_{20} reliability coefficient is shown in Equation 3.14. KR_{20} turns out to be an estimate of the average reliability coefficient for a test split in all possible ways, where the split-half psychometric reliability is calculated using the Rulon Equation. Although this equation is not shown here, due to an additional simplifying assumption, this method does not require the Spearman-Brown correction. As a result, KR_{20} results in a psychometric reliability coefficient that is known to be below the psychometric reliability values produced by split-half methods but to what extent is not known. This implies that KR_{20} reliability estimates a lower bound of the psychometric reliability for a test under development. However, despite the nuances, KR_{20} will always produce the same estimate of psychometric reliability for a given pool of questions and group of test takers.

Equation 3.14: Kuder-Richardson KR_{20}

$$KR_{20} = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum pq}{\sigma^2}\right)$$

where,

k is the number of questions
p equals the probability of answering a question correctly
q equals 1 - *p*
σ^2 is the variances of the total scores for all students

As is the case for coefficient alpha, estimation of the KR_{20} reliability coefficient that will result from a test comprised of separate sub-tests is possible using Equation 3.15. This equation is used to determine the total number of sub-tests, K, of equivalent length and having the same or psychometric reliability required for a certain whole test level of psychometric reliability. These concepts are illustrated in Example 3.5.

Equation 3.15: Prediction of KR_{20} from a Sub-test

$$Predicted\ KR_{20} = \frac{K \times KR_{20}}{1 + (K - 1) \times KR_{20}}$$

3.12 'if removed' Analysis

In Chapter 2, 'if removed' analysis was used to prune the item pool into the best possible combination of scales from which to build a sub-scale or summative scale. The same concepts will be used in this section to select the best combination of the questions not eliminated during Example 3.3. The goal is to produce a sub-test comprised of 5 questions.

The first step in 'if removed analysis', as applied to test development, is to eliminate one member for any pair of items that are perfectly correlated using the inter-correlation matrix shown in Table 3.10. As is the case when developing an observational instrument, the psychometric reliability is maximized when each item is only moderately correlated with any other item but strongly correlated with the total scores.

Using Equation 3.12, for 20 students involved in the comparison of each pair of items, d.f. = N = 40. For = 0.05, the size of required for significance equals 1.96 / (40 - 1)$^{0.5}$ = 0.314. Item 12 is both weakly or negatively correlated with most other items and item 3 is negatively with several question. These two questions will probably be eliminated from the sub-scale during pruning. Most other correlations are significant or nearly significant. Notice that items 2 and 10 are perfectly correlated. After reviewing the preliminary sub-scale, Question 10 is removed before proceeding with the 'if removed' analysis. We will include the remaining 9 questions in the 'If Removed Analysis' contained in Example 3.5.

Table 3.10: Inter-correlation Matrix for Items from Table 3.5

	2	3	4	6	8	9	10	11	12
1	0.89	0.28	0.56	0.52	0.30	0.57	0.89	0.42	0.10
2		0.34	0.43	0.44	0.38	0.64	**1.00**	0.49	0.15
3			-0.31	0.42	-0.24	-0.18	0.34	-0.21	0.33
4				0.31	0.30	0.57	0.43	0.42	-0.24
6					0.12	0.14	0.44	0.00	0.00
8						0.40	0.38	0.29	-0.19
9							0.64	0.49	-0.14
10								0.49	0.15
11									-0.17

Example 3.5

In the following example, a group of students from Example 3.4 were administered the preliminary sub-test comprised of the 9 question remaining after the inter-correlation analysis. Each question was scored 1 for a correct answer and 0 for an incorrect answer. The KR_{20} results for the initial version of this sub-test are shown in Equation 3.16.

In the last three rows of Table 3.11, p equals the proportion of correct responses for each question, r equals the correlation between the question scores and the total scores, once the effect of the question currently being correlated has been removed from the total scores. The last row contains the within question variance, p x q, for each question. For the initial test, the average proportion of correct responses equals 0.74 which is at the high end of the range of proportions of correct answers that results in a maximally discriminating test once guessing behavior is taken into consideration. The sum of the variances for each individual question equals 1.62 and the variance of the total scores equals 5.03. The KR_{20} results for the initial version of this sub-test are shown in Equation 3.16.

Equation 3.16: KR_{20} Psychometric Reliability

$$KR_{20} = \left(\frac{9}{9-1}\right)\left(1 - \frac{1.62}{5.03}\right) = 0.76$$

Table 3.11: Preliminary Sub-test

Ques.	**1**	**2**	**3**	**4**	**6**	**8**	**9**	**11**	**12**	Tot.	
Pat	1	1	1	1	1	1	1	1	1	9	
Rose	1	1	1	1	1	1	1	1	1	9	
Jill	1	1	1	1	1	1	1	1	1	9	
Jim	1	1	1	1	1	1	1	1	1	9	
Linda	1	1	1	1	1	1	1	1	1	9	
Joe	1	1	1	1	1	1	1	0	1	8	
Sam	1	1	1	1	1	1	1	1	0	8	
Don	1	1	1	1	0	1	1	1	1	8	
Nick	1	1	1	1	1	0	1	1	1	8	
Khoi	1	1	1	0	1	1	1	1	1	8	
Marty	1	1	1	0	0	1	1	1	1	7	
Hai	1	1	1	1	0	0	1	1	1	7	
Ray	1	1	0	1	0	1	1	1	1	7	
Lee	0	1	1	0	0	1	1	1	1	6	
Sam2	0	0	0	1	0	1	1	1	1	4	
Ty	0	0	0	1	0	1	1	1	0	3	
Patty	0	0	1	0	0	0	1	0	1	2	
Julie	0	0	1	0	0	1	0	0	1	2	
Pam	0	0	1	0	0	0	0	1	1	2	
Joey	0	0	1	0	1	0	0	0	1	2	
										5.03	$\mathbf{S^2}$
p	0.65	0.70	0.85	0.65	0.50	0.75	0.85	0.80	0.90		
r	0.91	0.93	0.20	0.55	0.51	0.40	0.65	0.48	0.05		
pq	0.23	0.21	0.13	0.23	0.25	0.19	0.13	0.16	0.09	1.62	$\sum$**pq**
										9	**k**
										0.76	$\boldsymbol{KR_{20}}$

At this point, 'if removed' analysis is used to prune the item pool. In this example, the psychometric reliability increases for all iterations when question are removed in the following order: 12, 3, 6 and 8. The sub-test is now comprised of questions 1, 2, 4, 9 and 11 as shown in Table 3.12. The KR_{20} value of 0.86 is in the range that would be expected.

Table 3.12: KR_{20} ' If Removed' Summary

Iter.	**1**	**2**	**3**	**4**	**6**	**8**	**9**	**11**	**12**
0	0.66	0.66	0.79	0.74	0.74	0.76	0.72	0.75	0.79
1	0.71	0.71	0.83	0.77	0.78	0.79	0.76	0.78	Rem
2	0.76	0.76	Rem	0.80	0.84	0.83	0.79	0.82	
3	0.79	0.78		0.82	Rem	0.85	0.80	0.83	
end	0.79	0.79		0.85		Rem	0.82	0.86	

The final question that must be answered is how many sub-tests of equivalent quality are needed to achieve a reasonable degree of psychometric reliability for the test as a whole. Using Equation 3.17, at test comprised of 5 sub-tests of equivalent quality will have a KR_{20} = 0.968. This is in the range typical for a relatively short standardized test. Resolution to the individual question level will likely not be possible but the correlations between sub-tests averages, for parallel forms of the test, will likely have adequate reliability.

Equation 3.14: Estimated KR_{20} for 5 Sub-tests

$$KR_{20} = \frac{5 \times 0.86}{1 + (5-1) \times 0.86} = 0.968$$

3.13 Concluding Remarks

The variation that is captured in any measurement contains components attributed to the response, which is the phenomena or behavior the measurement system is designed to measure and components attributed to a lack of accuracy and a lack of objectivity. All statistics meant to infer non-equivalence among measurements collected on a response, phenomena or behaviors, where an intervening event such as the passage of time or an experimental treatment has occurred, are required to truly measure the response, phenomena or behaviors of interest. Specifically, each measurement must contain 3% or less variation attributable to inaccuracy and 10% or less variation attributable to a lack of objectivity. If these thresholds are not rigorously adhered too, incorrect inferences can arise. This is serious because each incorrect inference has the potential to send the development and optimization efforts of an organization onto a path that leads to sub-optimal results.

Our ability to determine the reproducibility component of objectivity, for both tests and examinations, is impaired because students are irreversibly altered by exposure to a test or examination. This irreversible alteration casts doubt on the statistical independence of reproduced measurements.

The methods used to establish the quality of measurements resulting from measurements based around a standardized test as opposed to an instructor generated examination are somewhat different. This difference occurs because the three main sources of inaccuracy, content sampling, mis-classification and guessing, occur in different ways for these two types of assessment. These differences are not trivial and failure to acknowledge this fact, especially concerning question format, will slow the progress of the organization.

4

Minimum Correct Level of Performance

4.1 Introduction

In Chapter 3, we learned that controlling mis-classification errors, by minimizing the Type II error rate is critical for controlling the %inaccuracy to within the international limit of 3% or less within each measurement. The goal in this chapter is to demonstrate ways in which a minimum level of performance can be selected for instructor-generated examinations that allow the Type II error rate to be estimated. A requirement is that the method be a non-arbitrary method of selecting the minimum level of performance so that any person given the same information produces the same results. Our concern here is on examinations rather than standardized tests, because tests are developed in a controlled way that minimizes mis-classification errors.

Once a student has completed the instructional portion of an SOI, a measurement is collected in order to determine if the student has obtained adequate knowledge of the content of the SOI. What we really want to know is each student's true level of knowledge concerning the content. Unfortunately, the true level of knowledge is both unknown and unknowable. An examination score produces only an estimate of the true level of knowledge a student possesses, and this lack of a definitive determination sets up a classic situation in which a decision must be made, having only two decision states, when only incomplete information is available.

Any time a dichotomus decision is made with incomplete information there are not two, but four possible decision outcomes even though there are only two decision states. The first two decision outcomes are the most desirable and consist of a student being either correctly classified as possessing adequate knowledge of the content of an SOI or being correctly classified as possessing inadequate knowledge of the content of an SOI.

The third decision outcome occurs when a student who actually possesses an adequate level of knowledge of the content of an SOI but, for whatever reason, fails to demonstrate this fact during the measurement. When this student is not given credit for possessing the desired knowledge, a Type I error, having a probability of , has been committed. A Type I error has consequences for the student, but does not have serious consequences for the organization because the student does, in fact, possess the desired knowledge. A Type I error is easily corrected by simply reevaluating the student either in a similar manor, or using an alternative measurement strategy that allows the student to demonstrate his or her level of understanding. When the student is given credit for obtaining the knowledge presented in the SOI, the Type I error is corrected.

The fourth decision outcome, and the one of most interest, occurs when the score a student produces on an examination indicates that they have gained the knowledge presented in an SOI when in fact they actually possess an inadequate understanding of the content of the SOI. When the student is given credit for having gained the desired knowledge, a Type II error, having a probability of , has been committed. Type II errors are difficult to detect at the time of measurement and have serious implications for the student and for the organization. This is because a Type II error is typically discovered when the student produces actions contrary to those indicated by the instruction. Regarding the requirements for %Inaccuracy, only is added to the %

inaccuracy calculation. However, Type I errors can affect the costs of offering an SOI and so both are important.

4.2 The True Knowledge Distribution

Instructor-generated examinations, comprised of questions that are scored dichotomously, are the most common form of measurement in the fields of training and education. A method of modeling dichotomous data is to use the binomial distribution which was introduced in Chapter 3. Defining a binomial distribution requires two pieces of information. The first piece of information is the number of opportunities a student has to answer questions correctly, and this equals the number of questions on the examination. The second is the proportion of opportunities that are successful and this is equal to the proportion of questions answered correctly. This proportion is often simply called p. Determining the probability for a specific number of successes requires referring to the appropriate table of binomial probabilities.

For our purposes, we will use the binomial distribution to represent a hypothetical proxy student who's expected score equals p. The resultant distribution, describing the proxy students' knowledge, will be referred to as the True Knowledge distribution. This conception is known as the True Knowledge Model.

A key assumption of the True Knowledge model is that if all students are faithfully modeled by the true knowledge distribution, the relative frequency distribution of the student scores should match the frequencies of the binomial distribution used to model the proxy student.

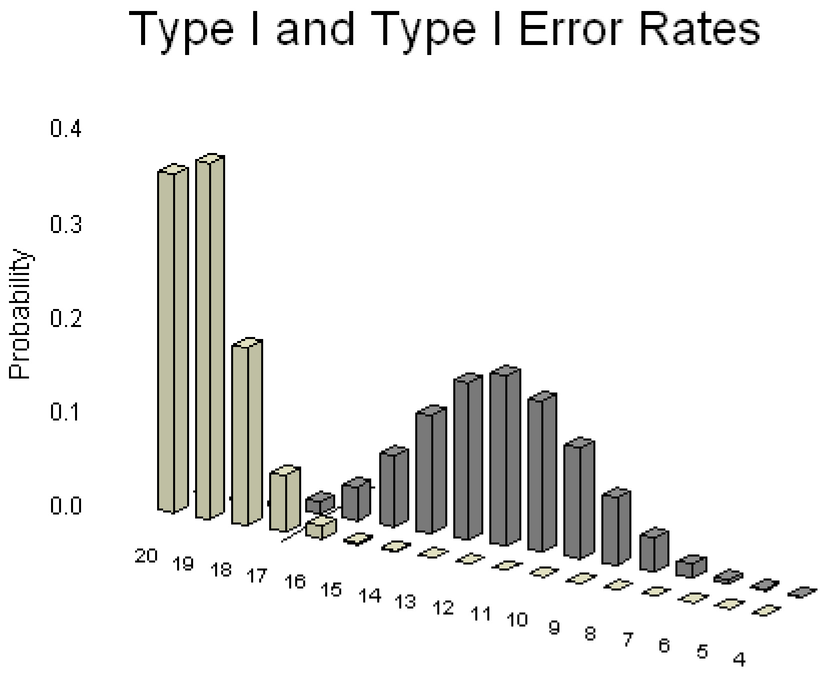
4.3 The Alternative Distribution

In order to estimate the minimum correct level of performance, abbreviated MC, a second binomial distribution is needed and this distribution is known as the alternative distribution. The alternative distribution represents a proxy student who has inadequate understanding of the content contained in an SOI. In order to define the alternative distribution, only the alternative distribution p is needed because the number of opportunities to answer a question correctly equals the number on the examination being modeled. The MC level of performance is completely determined by the alternative distribution. When no students are allowed to receive credit for obtaining the knowledge contained in an SOI who score below the MC level of performance on the examination for that SOI, the SOI is in statistical control.

The probability area bounded by a binomial distribution equals 1.0. The alternative distribution and the True Knowledge distribution overlap as shown in Figure 4.1. Figure 4.1 shows an example of this overlap region for a specific set of parameters. The value of must be less than or equal to the inaccuracy expectations for mis-classification errors. In this figure, the MC level of performance occur at 17 questions correct. Students scoring 16 questions correct or below are reassessed and then re-instructed if necessary.

The alternative distribution p can be selected by the organization, but is more commonly provided by an external source such as school officials or state legislators. In many cases, the goal of a certain amount of instruction is to produce a score above a predefined score on a standardized test such as a college entrance examination. The proportion of points desired on the standardized test becomes the alternative distribution p. Similarly, if a standardized test is given to students as a way of measuring their progress in K-12 education, the alternative distribution p equals the mandated minimum proportion of questions a student must answer correctly on the relevant portion of the Standardized Test. In certain cases, such as training designed to eliminate expensive or life threatening mistakes, there is no alternative distribution, because no level of performance less than a perfect score is acceptable. This case will be addressed later in this chapter.

Figure 4.1: In this figure, the line at 17 questions defines the and probability regions. The True Knowledge distribution p for this figure equals 0.949.



4.4 Selecting the Type I Error Rate

Selecting the target proportion of correct answers for a True Knowledge distribution depends upon the organizations expectations for the number of students that are re-assessed and re-instructed. Selecting a value for defines a specific True Knowledge distribution proportion of correct answers. What is a reasonable value for ?

If the baseline assumption that the value gained or costs avoided by providing an SOI is greater than or equal to the costs of providing the SOI, then the area contained in probability region should approximately equal the area contained in the probability region. In this case, the True Knowledge distribution p is determined by systematically varying the true knowledge proportion of correct answer until the ≈ . This solution is shown in Table 4.1, where the MC level of performance equals 15 questions correct. In this table, the alternative distribution p equals 0.60 and = 0.016. The closest value for ≈ occurs where the true knowledge proportion of correct answers equals p = 0.923. At this True Knowledge p, = 0.0158. When the True Knowledge proportion of correct answers is equal to the obtained score proportion of correct answers for the population and the shape of the obtained score distribution is similar to the True Knowledge distribution, the SOI is in statistical control as well as economic control. Economic control is critical for obtaining cost efficiency.

Other assumptions regarding the relative costs of offering an SOI compared to the benefits derived from or costs avoided by offering the SOI are possible. If the Type II error cost is large compared to the Type I error cost then ≈ / x is a reasonable assumption. The variable x is the ratio of the benefits or costs avoided to the costs of providing the instruction. Using this logic, if the cost of a Type II error is 3 times as large as the cost of resolving a Type I error, then economic control occurs were ≈ / 3.

Table 4.1: For N = 20, Alternative Distribution p = 0.60, ≤ MC = 15 questions

MC	**p = 0.6**	**T. K.**	**p = 0.923**
20	0.0000	20	1.0000
19	0.0000	19	0.7986
18	0.0005	18	0.4626
17	0.0036	17	0.1963
16	**0.0160**	16	0.0630
15	0.0510	**15**	**0.0158**
14	0.1256	14	0.0031
↓	↓	↓	↓
0	1.0000	0	0.0000

Having a small probability is not free. If an SOI is performing poorly, meaning that the right tail of the distribution of student scores systematically exceed the relative frequencies of the True Knowledge distribution, the Type I error rate will be relatively large. This in turn increases the number of students requiring re-testing and re-instruction and this in turn increases the total cost of providing the instruction. The best method of controlling the Type I error rate costs for a given alternative distribution is to implement only instruction that is highly effective.

4.5 Distributional Specifications

Distributional specifications are a way of forcing the left side of obtained score distribution for a group of students completing an SOI to match or preferably exceed the binomial distribution probabilities from the True Knowledge distribution. A minimum of two probabilities must be selected in order to constrain the obtained score distribution. These are typically the probability for missing only zero or one questions and another just above the MC level of performance. The goal when using only two probabilities for a distributional specification is to force the obtained score distribution to be unimodal distribution heavily weight to the left side.

Enforcing the True Knowledge distribution p from Table 4.1, which equals p =0.923, the top specification could be that more than 1 - 0.4626 = 0.5374 or ~ 54% of students should get 19 or more questions correct. The probability of getting 17 or more question correct equals 1 - 0.0630 = 0.9370 so, the bottom specification should be that ~ 94% of students should get 17 or more questions correct.

As mention earlier in this chapter, when the costs of mistakes are extremely high or loss of life can occur, no alternative distribution exists and so by default all students should obtain a perfect score on the examination for the SOI. This goal may be too strict for newly implemented SOI's but some objective method is needed for establishing when the efficacy of an SOI is questionable. In order to accomplish this, a level of True Knowledge proportion of correct answers that is very close to but less than 1.0 is selected. A true knowledge value used for this purpose is TK = 0.99. Table 4.2 below contains the binomial probabilities for a 20 question examination where the probability of a success equal p = 0.99

Table 4.2: TK = 0.99

Questions	**Prob.**	**Prob.**
Correct	**Individual**	**Cumulative**
20	0.818	1.000
19	0.165	0.182
18	0.016	0.017
17	0.001	0.001
16	0.000	0.000
15	0.000	0.000
14	0.000	0.000

If the relative frequency distribution of obtained scores exceeds the expected probabilities of a True knowledge distribution having p = 0.99 for 0 questions missed and is below the expected probabilities for the other numbers of questions missed, there is evidence that the obtained score distribution p is approaching p = 1.0. In order to constrain the obtained score distribution, two levels of correct question are again selected. The first is the minimum proportion of students who must obtain a perfect score on the examination which equals p = 0.818 or ~82%. The second value selected to constrain the distribution of obtained scores is that 1 - 0.017 = 0.983 meaning that 98.3% of the students must answer 18 or more questions correctly. If the distributional specifications are exceeded, a case can be made that the SOI is functioning to expectations.

4.6 Concluding Remarks

Selecting the minimum correct level of performance at the individual student level requires the presence of an identifiable alternative distribution. An alternative distribution is often imposed upon an SOI or collection of SOIs by a minimum external test score. One can also be selected by the organization. Regardless of the source, bringing an SOI into both statistical and economic control will require the average student to score well above what is now considered adequate.

When no alternative distribution exists, the assumption is that a perfect score on the examination is the only acceptable score at the individual student level. However, by selecting a True Knowledge distribution p very close to 1.0, of say 0.99, defines a binomial probability distribution. If the proportion of students missing zero questions exceeds those from the True Knowledge distribution and the proportion of students missing one or more questions is below the expected proportion, the SOI is in de facto control. De facto control is also a useful concept for selecting a performance criteria for externally developed instruction.

As we will see in Chapter 5, when a collection of SOIs, in which knowledge gained in one or more prior SOIs in the collection is necessary for success in the current SOI, high performance in each SOI is critical. If high attainment is not maintained, the cumulative Type II error rate becomes large.

5

Reliability

5.1 Introduction

For our purposes, the reliability of an SOI equals the probability that the results from each offering of the SOI exceeds some predefined performance limit. The performance limit is typically the MC level of performance when an alternative distribution exists. Distributional specifications are also used for purposes of establishing performance limits. Interestingly, the expected binomial proportions of a distributional specification can be used to create non-overlapping categories of performance into which each applicable obtained score becomes a member of only one category. Performance criteria that are not based upon statistical properties, such as the MC level of performance, the True Knowledge proportion of correct answer or upon distributional specifications, are often viewed as arbitrary by staff members. This situation should be avoided whenever possible.

At this point in the quality control implementation process, the measurement system for each SOI will have been defined and the MC level of performance and/or distributional specifications determined. The measurement system for each SOI will have been studied for accuracy and objectivity. Regardless of the measurement system configuration, the measurement system must contribute little variation to the resultant measurements. The value of reliability modeling is predicated upon having high quality measurements containing large amounts of extractable information.

SOIs may be similar to or related to other SOIs and for this reason collections of SOIs are often examined for reliability as either a category or a series as well as separately. For naming purposes, a collection of independent SOIs will be known as a category of SOIs and a collection of dependent SOIs will be known as a series of SOIs. Dependence exists when knowledge contained in one or more prior SOIs is required in order to perform satisfactorily during the current SOI. The relationship between an individual SOI in a category of SOIs is somewhat arbitrary and serves mainly as a management tool. The relationship between an individual SOI in a series of dependent SOIs is based on the content of each and the order in which each SOI is presented. For this reason, membership in a series of SOIs is not arbitrary and whenever a series exists, the reliability of the entire series, both for each individual SOI in the series and as well as for the series as a whole, must be determined and managed. A series of SOIs can be quite complex and in some cases, as with certain core competencies in a college curriculum, may extend over several semesters or even years.

The most common point in time where reliability is measured is directly after an SOI is completed and the reliability calculated from these measurements is known as the Initial Reliability. At any point in time after an SOI has been completed, often weeks or months later, the probability that an SOI is still functioning as desired can be determined and the reliability calculated from these measurements is known as the Transfer Reliability. It is reasonable to assume that instructional variables affect Transfer Reliability. In fact, certain instructional practices that lead to high Initial Reliability, such as excessive repetition, have a negative impact on Transfer Reliability. Determining the Transfer Reliability of each of the organization's SOIs, categories of SOIs and series of SOIs is possibly the most important and widely overlooked measurement of instructional effectiveness.

The most common forms of measurements regarding instructional outcomes are instructor generated examinations comprised of questions that are scored dichotomously. For the purposes of this chapter, we will confine our discussion to this type of measurement system. These concepts are easily adapted to other types of measurement systems if it is possible to identify a non-arbitrary score that represents the lower limit of acceptable performance.

5.2 The Failure Rate

When a student's score is below the minimum level of performance on the examination for a specific SOI, in reliability terminology this is known as a failure. The proportion of students' scoring below that level of performance equals the failure rate. Due to the fact that all instructional variables are under the control of the persons developing and implementing instances of instruction, a failure in this context refers to the failure of the SOI to perform as desired rather than the failure of the student per se.

Early in the implementation of an organization's quality control program, the failure rate for many already existing SOIs will be large. This is because, almost without exception, the reliability of pre-quality control processes will have been consistently and often dramatically over estimated. Accepting this fact is a make or break point during early implementation. When questions inevitably arise concerning what circumstances led to poor reliability, organizations having experience with quality control methodology, focus on determining the causes of low reliability and correcting these.

Discussions of factors beyond the control of the persons developing and implementing instruction are counterproductive for two reasons. First, the condition of the student population and their present circumstances are not likely to be changed, at least in the short-term, by instructors or the developers of instruction. These factors must be considered during development and implementation of instruction or improvements in reliability are unlikely. Second, assigning blame is counterproductive because regardless of the amount of blame assigned, at the end of this often divisive process, the problem of low reliability remains unaddressed. To restate, the organization has control over all of the variables that can result in highly reliable instruction and so systematic changes to these variables should be the area of focus for improvement efforts.

5.3 Reliability of Collections of SOIs

The knowledge that must be in place in order to perform successfully during the current SOI is known as antecedent knowledge. Independent SOIs as well as those contained in a series of SOIs can have antecedent knowledge requirements. The antecedent knowledge required for success in an independent SOI or for the first SOI in a series of SOIs is often acquired outside the organization. Regardless of the source of antecedent knowledge, the persons developing and implementing instruction are responsible for verifying and supplying antecedent knowledge if necessary.

As mentioned earlier, reliability modeling of a series of dependent SOIs requires establishing non-overlapping levels of performance. For purposes of explanation, the MC level of performance will be used in the following discussion concerning the development of non-overlapping performance categories.

Students' scoring at or above the MC level of performance enter the Above MC performance category. Students' not achieving the MC level of performance are reassessed and based upon these results are placed in one of two after-the-fact performance categories. These are the S+Above MC and the S+Below MC performance categories, where the S stands for reassessed. The S+Above MC performance category contains students' who demonstrate adequate knowledge upon further assessment and the S+Below MC performance category contains students' who, for whatever reason, fail to demonstrate adequate knowledge upon reassessment. The rationale for including after-the-fact performance categories is that becoming a member of the S+Above MC performance category corrects a Type I error. Membership in the S+Below MC performance category adds definiteness to the conclusion that the student did, in fact, not possess the desired knowledge.

Whenever after-the-fact performance categories or other modifications are used to improve the initial reliability, the reliability determined before the modification is known as the First Pass Initial Reliability. The reliability calculated from the combination of the Above MC and S+Above MC performance categories is known as the Initial Reliability.

Ideally and in practice, each SOI is required to demonstrate reliability before wide scale adoption. If the First Pass Initial Reliability or the Initial Reliability is dramatically below expectations, the SOI receives further development before another attempt is made to demonstrate efficacy.

After implementation, the primary strategy for increasing the reliability of an SOI is through a modification known as Redundancy. Redundancy is present if, when an SOI fails to achieve adequate reliability, the system has recourse to additional instruction on the topics presented during the SOI. Redundancy can also be embedded in subsequent SOIs. Elaborate systems of redundancy are accomplished through imbedding redundant portions of instruction within several subsequent SOIs in a series. The best examples of this are spiral curriculums common in grade K thru 12 education.

The reliability of imbedded redundancy is often difficult to determine and so imbedded redundancy is typically evaluated in isolation before incorporation into an SOI. When redundancy is studied separately, the reliability of the SOI equals the reliability of the SOI proper plus a component attributed to the reliability of the redundant instruction. Taken to extremes, if sufficient amounts of redundancy are provided, any SOI, category of SOIs or series of SOIs can be made exceedingly reliable. Of course, this strategy of achieving high reliability is limited by cost and negative effects on student morale and serves mainly as a point of discussion.

5.3 Reliability Models

There are two distinct models for evaluating reliability and selecting the correct one depends upon whether a category of SOIs or a series of SOIs is being modeled. Independence versus dependence is important because not only is this issue critical for selecting the correct reliability model, it is also critical for selecting the correct statistical analysis methods for comparing improvement efforts to the status quo. Situations of total dependence are less common than some degree of partial dependence. In situations of partial dependence, the methods used for reliability modeling and statistical analysis when dependence is present are the correct choices.

The model used for categories of SOIs is the simplest form and consists of multiplying all individual SOI reliability values. There is also a relatively simple method to determine the effects of redundancy in this model. The model that is provided for determining the reliability of series of SOIs is the Markov Chain model. This model is more complicated but has the advantage of being scalable to accommodate large series of SOIs. The method of determining the efficacy of redundant instruction is slightly more difficult than that used for categories of SOIs. These models are not interchangeable and using the correct one is critical. Our discussion begins with the reliability model appropriate for categories of SOIs.

5.5 Reliability of a Category of SOIs

When modeling independent SOIs, in the absence of redundancy, the Initial Reliability of the current SOI equals the proportion of students who score at or above the MC level of performance. It is often useful to group independent SOIs into categories of SOIs based upon some commonality among the individual SOIs. For example, all independent SOIs related to productivity enhancement could form one category and all independent SOIs related to safety could form a second category. Grouping independent SOIs into categories allows the SOIs to be managed both individually and as a group. In order to calculate the reliability of a category of SOIs, the initial reliability values for each individual SOI in the category, $P(R_x)$, are multiplied together as shown in Equation 5.1. This product equals the Category Reliability, R_c.

Equation 5.1: Category Reliability

$$R_c = P(R_1) \times P(R_2) \cdots \times P(R_x)$$

In many training situations, the majority of the courses are comprised of a single SOI. In the following example, an organization has four independent courses, each composed of a single SOI, and the desire is to manage these as a category. The ultimate desire is to avoid after-the-fact performance categories and other modifications where ever possible. The organization has determined that adequate performance is achieved when the category reliability equals $R_c \geq 0.80$ where all individual SOIs have an initial reliability value of $P(R_x) \geq 0.90$. The initial reliability values for each of the four SOIs equal 0.91, 0.98, 0.90 and 0.94 and so the category reliability equals $R_c = 0.754$ as shown in Equation 5.2. Notice that even though all of the individual SOIs have a initial reliability greater than or equal to 0.90, the R_c is well below 0.80.

Equation 5.2: Category Reliability for the Four SOIs

$$R_c = 0.91 \times 0.98 \times 0.90 \times 0.94 = 0.754$$

In order to reach the reliability target for the category, the third SOI is selected for improvement in the form of redundancy consisting of supplemental instruction for students scoring near to and below the MC level of performance. The extent of improvement to the first pass initial reliability of the third SOI is calculated using Equation 5.3 where $P(R_{xR})$ equals the initial reliability of the SOI due to the inclusion of the redundant instruction. In order to calculate the initial reliability for the third SOI, including the component attributed to the redundancy, all of the variables in Equation 5.3 must be defined.

Equation 5.3: Reliability of Redundancy

$$P\left(R_{xR}\right) = P\left(R_x\right) + \left(P\left(R_R\right) \times P\left(R_N\right)\right)$$

where,

$P(R_x)$ equals the reliability of the SOI without redundancy
$P(R_R)$ equals the reliability of the redundant portion of the SOI
$P(R_N)$ equals the probability of needing the redundant portion of the SOI

The first pass initial reliability for the third SOI equals $P(R_3) = 0.90$. If the bottom 12% of the students' are referred for additional instruction, then $P(R_N) = 0.12$. If 57% of students' score at or above the MC level of performance on a second but equivalent examination for the third SOI, $P(R_R) = 0.57$. Substituting these values into Equation 5.3, the initial reliability for the third SOI now equals $P(R_{3R}) = 0.90 + (0.57 \times 0.12) = 0.9684$. After improving the initial reliability of the third SOI, the category reliability equals $R_c = 0.91 \times 0.98 \times 0.9684 \times 0.94 = 0.812$ which is now above the desired value of $R_c \geq 0.80$.

What if, due to resource constraints, only a maximum of 7% of students can be given the redundant instruction? Using Equations 5.1 and 5.3, it is possible to estimate $P(R_{3R})$ and Rc if only 7% of students scoring below the MC level are given redundant instruction. In this case, we will assume that $P(R_R)$ still equals 0.57, for $P(R_N) = 0.07$, $P(R_{3R}) = 0.90 + (0.57 \times 0.07) = 0.94$. $R_c = 0.91 \times 0.98 \times 0.94 \times 0.94 = 0.79$ which is now slightly below the desired category reliability.

A major advantage of using reliability modeling is that by substituting hypothetical values for one or more of the actual values, it is possible to determine how much improvement is necessary to meet certain reliability expectations. For example , how much the first SOI must be improved in order to meet the category reliability target of $R_c \geq 0.80$ when $P(R_{3R}) = 0.94$? By substitution, the initial reliability of the first SOI must be greater than or equal to $P(R_1) = 0.924$ because $R_c = 0.924 \times 0.98 \times 0.94 \times 0.94 = 0.80$. What if upon on further inspection, the initial reliability for the second SOI is believed to be higher than would be expected at $P(R_2) = 0.98$. If a more realistic value of $P(R_2) = 0.95$ is substituted into the equation for R_c, then $P(R_1)$ must be greater than or equal to 0.954 in order to achieve $R_c = 0.954 \times 0.95 \times 0.94 \times 0.94 = 0.80$ for this category of SOIs.

In summary, grouping independent SOIs into categories based upon perceived similarities is useful for management purposes. A main benefit is that a category of SOIs can be modeled several ways in order to plan for adequate reliability. A more complex situation occurs when SOIs are dependent upon antecedent knowledge for successful performance. Our discussion now turns to this topic.

5.6 Reliability of Series of SOIs

If knowledge obtained during one or more prior SOIs influences the chances of success on the current SOI, the SOIs are dependent on prior performance. As mentioned earlier, a collection of dependent SOIs is known as a series of SOIs. When dependence is present, movements of students between levels of acceptable and unacceptable performance within the series are of interest in addition to the level of performance on the current SOI. Due to the fact that the pattern of student movements between performance categories is often unpredictable, a situation further complicated when considering a series of SOIs, there is no single equation that can be used to model this situation. There is however, a powerful model, called a Markov model, that is relatively easy to use and allows large numbers of performance categories and student centered variables to be included into a single reliability model. Markov models employ matrices and iterations of matrices to describe the movements of students between performance categories from one SOI to another. For our purposes, only the number iterations of a Markov model equal to the number of SOIs in the series are of interest.

A Markov model is comprised of two matrices, the Distribution Matrix and the Transition Matrix. The Distribution Matrix contains the number of students' belonging in each non-overlapping performance category before each iteration. The Transition Matrix contains the probabilities that students will move between the performance categories contained in the Distribution Matrix during a single iteration of the model. The output from an iteration provides the Distribution Matrix for the next iteration.

There are three specific requirements for the matrices used in a Markov model. First, the Transition Matrix must be a square matrix containing only non-negative transition probabilities and have the same number of columns as the Distribution Matrix. A square Transition Matrix has an equal number of rows and columns. Second, every transition probability must be between 0.0 and 1.0. Third, every row in a Transitions Matrix must sum 1.0.

Equation 5.4 shows a hypothetical Distribution Matrix D_0 and a Transition Matrix T_y. D_0 and T_y are multiplied, as shown in Equation 5.5, during each iteration of the Markov model. The product of Equation 5.5 equals D_1 where, D_1 become the input matrix for the next iteration of the model. The Transition Matrix may contain either the same probabilities for each iteration or unique probabilities reflecting differing conditions from one iteration to the next. This depends upon the goals of the model being developed.

Equation 5.4: The Markov Model

$$\textit{For } D_0 = \begin{bmatrix} a & b & c \end{bmatrix} \textit{ and } T_y = \begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{bmatrix}$$

Equation 5.5: Matrix Multiplication

$$D_0T_y = \begin{bmatrix} ax_1 + bx_2 + cx_3, & ay_1 + by_2 + cy_3, & az_1 + bz_2 + cz_3 \end{bmatrix} = D_1$$

5.7 Performance Categories

There are many ways to form non-overlapping performance categories for use in a Markov model. In fact, categories can be based upon any logical type of classification system. Two common systems are distributional specifications and the number of students performing above or below the MC level of performance. When an alternative distribution is available, the use of after-the-fact performance categories is also common.

When using after-the-fact performance categories, the first outcome occurs when students who scored at or above MC level of performance on the previous SOI in the series score at or above the MC level of performance on the current SOI. The second occurs when students who scored below the MC level of performance on the previous SOI in the series perform at or above the MC level of performance on the current

SOI. The third outcome occurs when students who scored at or above the MC level of performance on the previous SOI in the series score below the MC level of performance on the current SOI. The forth outcome occurs when students who scored below the MC level of performance on the previous SOI in the series continue to score below the MC level of performance on the current SOI.

More detailed outcomes and movements between outcomes, when an alternative distribution exists, are essentially subdivisions of one or more of the four outcomes described in the previous paragraph. Additional outcome categories are useful when students within one of these four outcome are believed to have different underlying probabilities of success on the current SOI. These can be either nominal or ordinal classifications. Two examples of nominal classifications are gender and ethnicity and two examples of ordinal classifications are prior education and job classification.

When using a distributional specification to define outcomes, performance categories are based upon the number of questions each student answers correctly. For example, on a 20 question examination, there are a distinct number of students who answer 20 questions correctly, a distinct number of the students who answer 19 questions correct and so on.

5.8 Forms Of the Markov Model

There are also two forms of Markov models: Markov chains and Markov processes. In a Markov chain, the transition probabilities remain fixed during each iteration of the model where as in a Markov process, the transition probabilities change in reaction to the preceding iteration.

For education and training applications being modeled for the first time, the Transition Matrix is populated by calculating the transition probabilities for each iteration of the model. This results in a Markov process model. A useful approach is to model a series of SOIs as a Markov process, then average the probabilities contained in the Transition Matrices for all iterations in order to develop a single Transition Matrix. This results in a Markov Chain model. This approach to modeling will be demonstrated in Example 5.2.

In the basic form of a Markov model, the Transition Matrix contains no row or rows having a transition probability equal to 1.0. When a transition probability equaling 1.0 exists, the resulting Markov model is said to contain an Absorbing State. When an Absorbing State is present, the resulting Model is now called an Absorbing Markov model. In an Absorbing outcome, due to the requirement that sum of each row of transition probabilities equals 1.0, all other transition probabilities in the row will equal 0.0 by definition. This creates a situation where once a student enters the Absorbing State they are unable to leave during subsequent iterations. A common example of an Absorbing State occurs when students are allowed to withdraw from a series of SOI before completion. When an Absorbing Markov model is iterated enough times, all of the students will eventually end up in the Absorbing State.

(Technical Detail): If there is a power to which the Transition Matrix for a Markov chain model can be raised that results in a Transition Matrix containing only transition probabilities greater than 0.0, the Transition matrix is known as a regular matrix. When the conditions for a regular matrix are met, a Markov Chain model will have a Steady State solution. The Steady State solution is reached when the Distribution Matrix and the output matrix for an iteration of the model are identical (End Detail).

For our purposes, only the first several iterations are of interest so, calculating a Steady State solution is typically omitted from the analysis. However, the Steady State solution for Example 5.1 is provided in order to illustrate this feature of a Markov Chain model.

Example 5.1

In this first example, a single course comprised of a series of three SOIs is modeled for a group of 500 students. In this example, it is possible to identify an alternative distribution and so performance categories based upon the MC level of performance are possible. This series of SOIs will be modeled as a Markov process thereby allowing the transition probabilities to change for each iteration of the model. There are three performance categories contained in the model. The first performance category is the Above MC performance category which contains students who correctly answer a number of questions greater than or equal to the MC level of performance on the examination for the current SOI.

The total number of students exiting the current SOI in the Above MC performance category, when divided by the total number of students, equals the First Pass Initial reliability for that SOI. A high First Pass Initial reliability is important because this is the only outcome that does not require additional resources above that already used to develop and deliver the instruction and assess students. As we will see in this example, the costs associated with additional assessments can become quite large.

The second performance category, the S+Above MC performance category, contains students who initially score below the MC level of performance on the examination for the current SOI but then demonstrate adequate knowledge during a subsequent 'screening' assessment. The total number of students exiting an SOI in both the Above MC and the S+Above MC performance categories when divided by the total number of students equals the Initial Reliability for the current SOI.

The third performance category, the S+ Below MC performance category, contains students who score below the MC level of performance and do not demonstrate adequate understanding upon subsequent reassessment. As mentioned in the introduction, a main rationale for including after-the-fact performance categories is that the S+Above MC performance category allows us to correct Type I errors by providing another assessment opportunity for lower performing students.

The assumption in this example is that all students entering the first SOI have adequate antecedent knowledge as evidenced by prior performance. In the absence of any information to the contrary, all students are then placed in the Above MC performance category in the beginning Distribution Matrix, D_0, as shown in Equation 5.6. A further assumption made in this example is that students are not allowed to withdraw from the series so the model does not contain an Absorbing State.

Equation 5.6: Initial Distribution Matrix

$$D_0 = \begin{bmatrix} 500 & 0 & 0 \end{bmatrix}$$

On the examination for the first SOI, 430 students score Above MC level of performance and so the First Pass Initial Reliability equals 430/500 = 0.86. Of the 70 students who scored below the MC level of performance, 40 / 500 or 8% of the students enter the S+Above MC performance category upon reassessment and 30 / 500 or 6% of the students enter the S+Below MC performance category upon reassessment. If a high proportion of the students requiring additional assessment enter the S+Above MC performance category, the measurement system should be re-evaluated. The Initial Reliability for the first SOI in this series equals (430 + 40) / 500 = 0.94. The results of the first iteration are shown in Equation 5.7 where the output matrix from the first iteration, D_1 becomes the Distribution Matrix for the second iteration.

Equation 5.7: First Iteration

$$D_1 = \begin{bmatrix} 500 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0.86 & 0.08 & 0.06 \\ 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 \end{bmatrix} = \begin{bmatrix} 430 & 40 & 30 \end{bmatrix}$$

During the second SOI, of the 430 students who entered in the Above MC performance category, 396 of 430 or 92% remain in the Above MC performance category, 17 of 430 or 4% move into the S+Above MC performance category and 17 of 430 or 4% move into the S+Below performance category. Of the 40 students who enter the second SOI in the S+Above MC performance category, 33 of 40 or 82% remain in the S+Above MC performance category while 3 of 40 or 8% enter the Above MC performance category and 4 of 40 or 10% enter the S+Below MC performance category. Of the 30 students who enter the second SOI in the S+Below MC performance category, 19 of 30 or 64% remain in the S+Below performance category while 1 of 30 or 4% enters the Above MC performance category and 10 of 30 or 33% enter the S+Above MC performance category. The second iteration of the model is shown in Equation 5.8 where the First Pass Initial reliability equals (396 + 3 + 1) / 500 = 400 / 500 = 0.80 and the Initial reliability for second SOI equals (400 + 60) / 500 = 0.92. Unfortunately, 60 + 40 = 100 reassessments are required to achieve these results.

Equation 5.8: Second Iteration

$$D_2 = \begin{bmatrix} 430 & 40 & 30 \end{bmatrix} \begin{bmatrix} 0.92 & 0.04 & 0.04 \\ 0.08 & 0.82 & 0.10 \\ 0.04 & 0.32 & 0.64 \end{bmatrix} = \begin{bmatrix} 400 & 60 & 40 \end{bmatrix}$$

Entering the third SOI, there are now 400 students in the Above MC performance category, 60 students in the S+Above MC performance category and 40 students in the S+Below MC performance category. The Transition probabilities for the changes between the performance categories are calculated in the same manner as that used for the second SOI and the results are shown in Equation 5.9. Notice that after the third iteration, no students who entered the third SOI in the Above MC performance category completes this SOI in the S+Below MC performance category. Conversely, no students who enter the third SOI in the S+Below MC performance category completes this SOI in the Above MC performance category. Here again, even though the vast majority of the students who entered the third SOI in the Above MC remained in this category, the First Pass Initial reliability has declined to 359 / 500 = 0.72. The Initial Reliability for the third SOI has increased slightly to (359 + 103) / 500 = 0.94. Unfortunately, 103 + 38 = 141 reassessments were required in order to achieve this result. It is clear that High First Pass Initial reliability is a requirement for economically viable instruction.

Equation 5.9: Third Iteration

$$D_3 = \begin{bmatrix} 400 & 60 & 40 \end{bmatrix} \begin{bmatrix} 0.88 & 0.12 & 0.00 \\ 0.12 & 0.80 & 0.08 \\ 0.00 & 0.18 & 0.82 \end{bmatrix} = \begin{bmatrix} 359 & 103 & 38 \end{bmatrix}$$

In this example, even though more than 90% of the students completed each SOI in either the Above MC or S+Above MC performance categories, a grand total of 40 + 30 + 60 + 40 + 103 + 38 = 311 reassessments were required, a situation that would make the cost of operating this series of SOI prohibitive. Better assessment may help to differentiate between students in the three performance categories but improving First Pass Initial reliability will be required to bring the total number of additional assessments into control. One way to achieve this is through the use of redundant instruction aimed at minimizing the number of students who enter the S+Above MC and S+Below MC performance categories.

If a decision is made to provide redundant instruction, the question becomes where to insert redundancy in order to produce the largest beneficial effect upon First Pass Initial reliability. One complication is that because multiple outcomes are possible for each student during each SOI, no single equation exists for estimating the effects of redundancy upon the movements of students between performance categories. When dealing with dependent SOI, it is difficult to select the place or places to insert redundancy that will have the greatest positive impact upon performance in the absence of reliability modeling. Initially, each instance of redundancy should be modeled as a separate SOI so that the effectiveness of the redundancy can be determined unambiguously. Once the efficacy of the redundancy is established, the lessons learned from the redundancy are incorporated into the SOI that caused the need for the redundancy. In Example 5.2, redundancy is inserted after the second SOI in order to bolster the reliability of the third SOI. This redundancy is treated as a separate SOI in this example.

Example 5.2

In Example 5.1, the Initial reliability continues to degrade during the second SOI so, rather than proceeding to the third SOI, redundancy is inserted between the second and third SOI. The redundancy consists of additional instruction, covering all topics contained in the first and second SOI, for all students exiting the second SOI in either the S+Above MC and the S+Below MC performance categories. The goal is to have no students remain in the S+Above MC performance category after the redundant instruction. In essence, all possible ambiguities concerning the performance of these students will have been resolved before these students proceed to the third SOI in the series.

Equation: 5.10: Equation 5.8 from Example 1plus Redundancy

$$D_2 = \begin{bmatrix} 430 & 40 & 30 \end{bmatrix} \begin{bmatrix} 0.92 & 0.04 & 0.04 \\ 0.08 & 0.82 & 0.10 \\ 0.04 & 0.32 & 0.64 \end{bmatrix} = \begin{bmatrix} 400 & 60 & 40 \end{bmatrix}$$

$$D_{R \to 3} = \begin{bmatrix} 0 & 60 & 40 \end{bmatrix} \begin{bmatrix} 1.00 & 0.00 & 0.00 \\ 0.72 & 0.00 & 0.28 \\ 0.25 & 0.00 & 0.75 \end{bmatrix} = \begin{bmatrix} 53 & 0 & 47 \end{bmatrix}$$

$$D_{3R} = \begin{bmatrix} 453 & 0 & 47 \end{bmatrix} \begin{bmatrix} 0.88 & 0.12 & 0.00 \\ 0.12 & 0.80 & 0.08 \\ 0.00 & 0.18 & 0.82 \end{bmatrix} = \begin{bmatrix} 399 & 63 & 38 \end{bmatrix}$$

As shown in iteration $D_{R\ ->3}$, of the 60 students exiting the second SOI in the S+Above MC performance category, 43 or 72% entered the Above MC performance category as a result of the redundant instruction. However, 17 will enter the third iteration in the S+below MC performance category. Of the 40 students exiting the second SOI in the S+Below MC performance category 10 or 25% entered the Above MC performance category. The total number of students entering the third SOI in the Above MC performance category now equals 400 + 53 = 453. The number of students entering the third SOI in the S+ Below MC performance category equals 30 + 17 = 47. The Initial Reliability for the second SOI in combination with the redundancy now equals 453 / 500 = 0.91 which is slightly below the reliability of the second SOI without the redundant SOI. The difference represents the resolution of Type II error occurring in the S+Above MC performance category. This is important because the screening procedure is meant to clarify mis-classification errors not create additional TypeII errors.

However, during the third SOI, if we assume that the transition probabilities are identical to those existing before the redundancy for the second SOI, the First Pass Initial reliability becomes 399 / 500 = 0.80, a distinct improvement over the 359 / 500 = 0.72 resulting without redundancy. The Initial Reliability becomes 462 / 500 = 0.92 which is equal to the results obtained for the second SOI without redundancy. Due to the fact that 40 additional students exited the third SOI in the Above MC performance category and 40 fewer students exited in the S+above MC performance category, there is a benefit to providing this redundancy.

Would the application of equally effective redundancy provide more benefit if inserted after the first SOI? In order to answer this question, the Transition Matrix for the redundant instruction is inserted between the first and second SOI where the students exiting the first SOI in the S+Above MC and S+Below MC performance categories become the Distribution Matrix for a redundant SOI as shown in Equation 5.11.

Equation 5.11: Equation 5.7 from Example 1plus Redundancy

$$D_1 = \begin{bmatrix} 500 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0.86 & 0.08 & 0.06 \\ 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 \end{bmatrix} = \begin{bmatrix} 430 & 40 & 30 \end{bmatrix}$$

$$D_{R\rightarrow 2} = \begin{bmatrix} 0 & 40 & 30 \end{bmatrix} \begin{bmatrix} 1.00 & 0.00 & 0.00 \\ 0.72 & 0.00 & 0.28 \\ 0.25 & 0.00 & 0.75 \end{bmatrix} = \begin{bmatrix} 36 & 0 & 34 \end{bmatrix}$$

$$D_{2R} = \begin{bmatrix} 466 & 0 & 34 \end{bmatrix} \begin{bmatrix} 0.92 & 0.04 & 0.04 \\ 0.08 & 0.82 & 0.10 \\ 0.04 & 0.32 & 0.64 \end{bmatrix} = \begin{bmatrix} 430 & 29 & 40 \end{bmatrix}$$

$$D_3 = \begin{bmatrix} 430 & 29 & 40 \end{bmatrix} \begin{bmatrix} 0.88 & 0.12 & 0.00 \\ 0.12 & 0.80 & 0.08 \\ 0.00 & 0.18 & 0.82 \end{bmatrix} = \begin{bmatrix} 382 & 82 & 35 \end{bmatrix}$$

When equally effective redundancy is inserted between the first and second SOI in this series, the First Pass Initial reliability of the second SOI equals 430 / 500 = 0.86 and the Initial Reliability of the second SOI equals 459 / 500 = 0.92. Both of these values are substantially above the values occurring before the application of redundancy. The First Pass Initial Reliability of the third SOI equals 382 / 500 = 0.76, which is again an improvement over not having redundancy but is slightly lower than that achieved by inserting the redundancy between the second and third SOI. The Initial Reliability equals 464 / 500 = 0.93 which is only slightly higher for this series of SOI without redundancy or when redundancy is inserted between the second and third SOI.

In either redundancy scenario, the number of additional assessments are still prohibitively large. Dramatic improvements in the First Pass Initial Reliability for each SOI in this series will be required in order to limit the number of additional assessments. For instance, if the First Pass Initial Reliability for each SOI equals 0.98, if the transition probabilities are still valid, then the maximum number of additional assessments would be no larger than 30.

5.9 Markov Chain Model

It is sometime advantageous to convert a Markov Process into a Markov Chain for modeling purposes. In order to accomplish this, the Transition probabilities from each of the Transition Matrices in a Markov Process model are averaged. The Average Transition Matrix is then used to model the series of SOI as a Markov Chain. When calculating the Transition probabilities for an Average Transition Matrix, the transition probabilities for outcomes for which membership is not possible are ignored when averaging. This occurs when D_0 contains categories having no members. An example of this occurs in Equation 5.6 where all of the students entering the first SOI are originally placed in the Above MC performance category in the initial distribution matrix D_0. This configuration precludes transition probabilities for the S+Above MC and S+Below MC performance categories during the first iteration. Using a Markov Chain simplifies Interpretation of the results from modeling because of the fixed Transition probabilities. Modeling a hypothetical scenario using a Markov Chain model is illustrated in Example 5.3.

Example 5.3

Using the information from Example 5.1, Equations 5.12 and 5.13 show the results produced by modeling the three SOI contained in this Example 5.1 as a Markov Chain. Notice that the compositions of the Distribution Matrices for the three iterations are approximately equal to those resulting from modeling these three SOI as a Markov Process.

Equation 5.12: Average Transition Matrix

$$D_{1_{Avg}} = \begin{bmatrix} 500 & 00 & 00 \end{bmatrix} \begin{bmatrix} 0.89 & 0.08 & 0.03 \\ 0.10 & 0.81 & 0.09 \\ 0.02 & 0.25 & 0.73 \end{bmatrix} = \begin{bmatrix} 445 & 40 & 15 \end{bmatrix}$$

where,

$$T_{Avg} = \begin{bmatrix} \dfrac{(0.86+0.92+0.88)}{3} & \dfrac{(0.08+0.04+0.12)}{3} & \dfrac{(0.06+0.04+0.00)}{3} \\ \dfrac{(0.08+0.12)}{2} & \dfrac{(0.82+0.80)}{2} & \dfrac{(0.10+0.08)}{2} \\ \dfrac{(0.04+0.00)}{2} & \dfrac{(0.32+0.18)}{2} & \dfrac{(0.64+0.82)}{2} \end{bmatrix}$$

Equation 5.13: Iterations 2 and 3 Using Average Transition Matrix

$$D_{2\,Avg} = \begin{bmatrix} 445 & 40 & 15 \end{bmatrix} \begin{bmatrix} 0.89 & 0.08 & 0.03 \\ 0.10 & 0.81 & 0.09 \\ 0.02 & 0.25 & 0.73 \end{bmatrix} = \begin{bmatrix} 400 & 72 & 28 \end{bmatrix}$$

$$D_{3\,Avg} = \begin{bmatrix} 400 & 72 & 28 \end{bmatrix} \begin{bmatrix} 0.89 & 0.08 & 0.03 \\ 0.10 & 0.81 & 0.09 \\ 0.02 & 0.25 & 0.73 \end{bmatrix} = \begin{bmatrix} 364 & 97 & 39 \end{bmatrix}$$

As mentioned in the introduction to this chapter, in instructional settings most interest is focused upon a finite number of iterations as a result of having a finite number of SOI. However, occasionally the equilibrium point, called the Steady State solution, is of interest. When a Transition Matrix is multiplied by an identical Transition Matrix, the resulting matrix is the square of that Transition Matrix. If there exists a power to which a Transition Matrix can be raised that produces a matrix that contains only transition probabilities greater than zero, a Markov Chain model based upon this Transition Matrix will have a Steady State solution if iterated enough times. By definition, Absorbing States contain transition probabilities of 0.0's and a 1.0 so no meaningful Steady State exists. The Steady State solution is reached when the Distribution Matrix for an iteration is identical to the output matrix. When the situation being modeled consists of a series of SOI for which there are no absorbing performance categories, there is a good chance that all of the transition probabilities will be greater than zero and so a Steady State solution will exist. In Example 5.3, 23 iterations are required to obtain the Steady State solution as shown in Equation 5.14.

Equation 5.14: Steady State Solution

$$D_{23\,Avg} = \begin{bmatrix} 206 & 204 & 90 \end{bmatrix} \begin{bmatrix} 0.89 & 0.08 & 0.03 \\ 0.10 & 0.81 & 0.09 \\ 0.02 & 0.25 & 0.73 \end{bmatrix} = \begin{bmatrix} 206 & 204 & 90 \end{bmatrix}$$

5.10 Absorbing States

If there exists a sequence of movements between performance categories for which the population being modeled can enter but not exit, the model contains an Absorbing State. An Absorbing State can be recognized because the row containing the transition probabilities will contain a single probability of 1.0 with the remaining probabilities equal to 0.0. As mentioned earlier, the most common example of a absorbing performance category occurs when students are allowed to withdraw from a particular series SOIs before completion. There is another type of Absorbing State that is unique to instructional processes. This is known as a de facto Absorbing State. A de facto Absorbing State occurs when students are allowed to slip into a pattern of hopelessly low performance for which the organization provides no recourse. This pattern is typically a result of the failure of one or more of the SOIs in a series. A de facto Absorbing State can also be caused thru systemic de-motivation of students resulting from prolonged exposure to a spiral curriculum. De facto Absorbing States are common in lengthy series of SOI such as those created under a spiral curriculum unless the Type II error rate for each SOI in the series is kept small. In Example 5.4, a hypothetical school district, employing a spiral curriculum, is modeled as an Absorbing Markov Chain model having a de facto Absorbing State.

Example 5.4

In this example, Grades 7 thru 12, for a hypothetical school district, are modeled using an Absorbing Markov Chain model. There are four performance categories: Green, Blue, Orange and Red. Green indicates the highest incoming performance and Red the lowest. The Red performance category is a de facto Absorbing State and students who enter the Red performance category will ultimately withdraw before completing Grade 12 even though they may continue attending school for years after entering the Red performance category. In Equation 5.15, notice that the bottom row of the Transition Matrix, describing the movements within the Red performance category, contains a probability equal to 1.0 and so by definition, the probability of leaving this performance category during subsequent grades equals 0.0.

In this hypothetical school district, of the 1,000 students entering Grade 7, 8.1% and 59.4% enter Grade 7 in the Green and Blue performance categories, respectively. Of the remaining students, 26.5% enter as members of the Orange performance category. The orange performance category has the most communication with the Red performance category which initially contains 6% of the students. The Transition probabilities for this Absorbing Markov Chain model were derived from an analysis of the school district’s prior performance. The Grade 7 thru 12 iterations are shown in Equation 5.15.

Upon inspection of the Grade 12 Distribution Matrix, notice that even though only 0.01 x 100 = 1% of the students in the Blue performance category and only 0.05 x 100 = 5% of the students in the Orange performance category enter the Red performance category during each iteration, by Grade 12, (179 / 1000) x 100 = 17.9% of the students enter Grade 12 as member of the Red performance category. These students will either have already withdrawn from school or be destined to withdraw during Grade 12. All totaled, 20.4% of students ultimately enter the Red performance category by the end of Grade 12. Notice that of the remaining students, 42.7% complete the Grade 12 in the Orange performance category. This is problematic because most of the Type II errors occur in the Orange performance category. Having such a large number of students completing Grade 12 in the Orange performance category virtually insures that a high Type II error rate exists.

Equation 5.15: Grade 7 thru 12

$$D_{Grade7}\begin{bmatrix}81 & 594 & 265 & 60\end{bmatrix}\begin{bmatrix}0.73 & 0.26 & 0.01 & 0.00\\ 0.04 & 0.65 & 0.30 & 0.01\\ 0.02 & 0.16 & 0.77 & 0.05\\ 0.00 & 0.00 & 0.00 & 1.00\end{bmatrix}=\begin{bmatrix}88 & 450 & 383 & 79\end{bmatrix}$$

$$D_{Grade8}\begin{bmatrix}88 & 450 & 383 & 79\end{bmatrix}\begin{bmatrix}0.73 & 0.26 & 0.01 & 0.00\\ 0.04 & 0.65 & 0.30 & 0.01\\ 0.02 & 0.16 & 0.77 & 0.05\\ 0.00 & 0.00 & 0.00 & 1.00\end{bmatrix}=\begin{bmatrix}90 & 376 & 431 & 103\end{bmatrix}$$

$$D_{Grade9}\begin{bmatrix}90 & 376 & 431 & 103\end{bmatrix}\begin{bmatrix}0.73 & 0.26 & 0.01 & 0.00\\ 0.04 & 0.65 & 0.30 & 0.01\\ 0.02 & 0.16 & 0.77 & 0.05\\ 0.00 & 0.00 & 0.00 & 1.00\end{bmatrix}=\begin{bmatrix}89 & 337 & 445 & 128\end{bmatrix}$$

$$D_{Grade10}\begin{bmatrix}89 & 337 & 445 & 128\end{bmatrix}\begin{bmatrix}0.73 & 0.26 & 0.01 & 0.00\\ 0.04 & 0.65 & 0.30 & 0.01\\ 0.02 & 0.16 & 0.77 & 0.05\\ 0.00 & 0.00 & 0.00 & 1.00\end{bmatrix}=\begin{bmatrix}88 & 314 & 445 & 154\end{bmatrix}$$

$$D_{Grade11}\begin{bmatrix}88 & 314 & 445 & 154\end{bmatrix}\begin{bmatrix}0.73 & 0.26 & 0.01 & 0.00\\ 0.04 & 0.65 & 0.30 & 0.01\\ 0.02 & 0.16 & 0.77 & 0.05\\ 0.00 & 0.00 & 0.00 & 1.00\end{bmatrix}=\begin{bmatrix}85 & 298 & 438 & 179\end{bmatrix}$$

$$D_{Grade12}\begin{bmatrix}85 & 298 & 438 & 179\end{bmatrix}\begin{bmatrix}0.73 & 0.26 & 0.01 & 0.00\\ 0.04 & 0.65 & 0.30 & 0.01\\ 0.02 & 0.16 & 0.77 & 0.05\\ 0.00 & 0.00 & 0.00 & 1.00\end{bmatrix}=\begin{bmatrix}83 & 286 & 427 & 204\end{bmatrix}$$

After considerable discussion among the staff concerning the unacceptably poor reliability of Grades 7 through 12, two broad remediation strategies emerge, only one of which will be developed and implemented. In order to determine which strategy is superior, the two remediation strategies are modeled for Grades 7 through 12 using the Transition Matrix from Example 5.4. The first remediation strategy, $D_{Grade12A}$, proposes to raise the performance of the students entering Grade 7 to a very high level by focusing upon improving the reliability of Grades 1 through 6. In this strategy, it is hoped that this added preparation will have a positive effect on subsequent performance. The Distribution Matrix for the Grade 7 iteration becomes $D_{Grade\ 7}$ = [950 50 0 0].

The Grade 12 results for the first remediation strategy are shown in Equation 5.16. This strategy has the prospect of reducing the number of students who ultimately enter the Red performance category before completing Grade 12 to 6.6% which is a vast improvement over the 20.4% occurring in Example 5.4. Unfortunately, a large number students, 37.1%, still complete Grade 12 in the orange performance category which again risks having a large Type II error rate.

Equation 5.16: Remediation Strategy Grade 12 A

$$D_{Gradel2A} = \begin{bmatrix} 246 & 377 & 331 & 45 \end{bmatrix} \begin{bmatrix} 0.73 & 0.26 & 0.01 & 0.00 \\ 0.04 & 0.65 & 0.30 & 0.01 \\ 0.02 & 0.16 & 0.77 & 0.05 \\ 0.00 & 0.00 & 0.00 & 1.00 \end{bmatrix} = \begin{bmatrix} 202 & 362 & 371 & 66 \end{bmatrix}$$

The second remediation strategy, $D_{Grade12B}$, proposes to raise the reliability of Grades 7 through 12 to a very high level while not altering the reliability of instruction students receive prior to Grade 7. The Transition Matrix for the Grade 12 iteration of this remediation strategy is shown in Equation 5.17. This Transition Matrix has the effect of restricting communication between the Green and Blue performance categories with the Orange performance category. As a result, even though 11.5% of the students ultimately enter the Red performance category and withdraw before completing Grade 12, only 2.5 percent of the students complete Grade 12 in the Orange performance category. This remediation strategy modestly increases the final number of students exiting Grade12 in the Red performance category, when compared to Equation 5.16, but this increase could simply represent a resolution of Type II errors. Due to the relatively small number of students exiting Grade 12 in the Orange performance category, this solution is likely to result in a relatively small Type II error rate.

Equation 5.17: Remediation Strategy Grade 12B

$$D_{Grade12B} = \begin{bmatrix} 100 & 769 & 25 & 106 \end{bmatrix} \begin{bmatrix} 0.95 & 0.03 & 0.01 & 0.01 \\ 0.01 & 0.95 & 0.03 & 0.01 \\ 0.01 & 0.95 & 0.03 & 0.01 \\ 0.00 & 0.00 & 0.00 & 1.00 \end{bmatrix} = \begin{bmatrix} 103 & 757 & 25 & 115 \end{bmatrix}$$

5.11 Transfer Reliability

In the introduction to this chapter, it was mentioned that the reliability of an SOI can and should be measured at one and possibly more than one points in the future, often long after the SOI has been completed. An example of a time period appropriate for safety training is 364 days. The goal, in this case, is to determine if the training is still functioning before an annual refresher. Another example is a 90 day period for students in K thru 12, which aims to insure that knowledge endures over the summer vacation. When Markov reliability analysis is applied to measurements collected to assess the amount of knowledge transferred by a student into long term memory in a way that allows for retrieval, the result is known as the Transfer Reliability. The goals of instruction are always long-term retention of the knowledge contained in the instruction and for this reason Transfer Reliability is possibly the most important and widely overlooked measure of instructional effectiveness. Achieving widespread and high Transfer Reliability is largely uncharted territory. However, it is known that even though Transfer Reliability is correlated with Initial Reliability, many strategies that lead to high Initial Reliability, such as excessive repetition, have a negative impact on Transfer Reliability. Ultimately, the goal is to have the Transfer Reliability equal to or possibly even higher than the Initial Reliability.

The Transfer Reliability for Example 5.2 is modeled in the same way as that used to model Initial Reliability. The assessment is an instructor generated examination prepared using the methods presented in Chapter 3 for producing parallel forms of an examination. There are two aspects of Transfer Reliability that are of most interest. The first concerns the Transfer Reliability of each SOI in a series modeled individually as a single iteration of a Markov model. The second concerns the movements of students between the performance categories during each iteration. In the following example, both aspects of Transfer Reliability are modeled.

Example 5.5

In the first part of this example, the three dependent SOIs from Example 5.2 are assessed for Transfer Reliability where each SOI is modeled individually. In Example 5.2, the S+Above MC and the S+Below MC performance categories are determined after-the-fact but in this example of Transfer Reliability calculations, no attempt is made to resolve mis-classification errors using a reassessment. Initial Transfer reliability is of most interest but it is conceivable that after-the-fact performance categories could be added. In this example, the output Matrix from each iteration becomes the Distribution Matrix for the Transfer Reliability model for that particular iteration. Due to the fact that no after-the-fact performance categories are considered in this example, members of the S+Above MC performance category enter either the Above MC or Below MC performances categories during each iteration of the Transfer Reliability model. Equation 5.17 to 5.19 contain the Transfer Reliability calculations for each SOI contained in Example 5.2.

Equation 5.18: Independent Transfer Reliability for Iteration 1

$$D_{1T} = \begin{bmatrix} 430 & 40 & 30 \end{bmatrix} \begin{bmatrix} 0.62 & 0.00 & 0.38 \\ 0.32 & 0.00 & 0.68 \\ 0.08 & 0.00 & 0.92 \end{bmatrix} = \begin{bmatrix} 282 & 0 & 218 \end{bmatrix}$$

Equation 5.19: Independent Transfer Reliability for Iteration 2

$$D_{2T} = \begin{bmatrix} 400 & 60 & 40 \end{bmatrix} \begin{bmatrix} 0.66 & 0.00 & 0.34 \\ 0.76 & 0.00 & 0.24 \\ 0.06 & 0.00 & 0.94 \end{bmatrix} = \begin{bmatrix} 312 & 0 & 188 \end{bmatrix}$$

Equation 5.20: Independent Transfer Reliability for Iteration 3

$$D_{3T} = \begin{bmatrix} 359 & 103 & 38 \end{bmatrix} \begin{bmatrix} 0.48 & 0.00 & 0.52 \\ 0.19 & 0.00 & 0.81 \\ 0.02 & 0.00 & 0.98 \end{bmatrix} = \begin{bmatrix} 193 & 0 & 307 \end{bmatrix}$$

When each SOI in the series is modeled independently, the Initial Transfer reliability of the first SOI equals 282 / 500 = 0.564 and for the second SOI equals 312 / 500 = 0.624. By the third SOI, the Transfer reliability equals 193 / 500 = 0.388. Notice that for the second SOI, 76% of the students exiting this SOI in the S+Above MC performance category enter the Above MC performance category when assessed for Transfer Reliability. For these students, their somewhat uncertain initial knowledge of the content of this SOI becomes more retrievable with the passage of time when compared to the results for the first and third SOI.

In the second part of this example, the Transfer Reliability for these SOIs are evaluated for Transfer Reliability as a series using a Markov Process model. This modeling approach produces the same Transfer Reliability results as those obtained when modeling each SOI in isolation but now the movements between performance categories within this series are available for evaluation. In this model, the Distribution and Transition Matrices for the first iteration are identical to Equation 5.18.

Equation 5.21: Transfer Reliability Modeled as a Markov Process

$$D_{1r*} = \begin{bmatrix} 430 & 40 & 30 \end{bmatrix} \begin{bmatrix} 0.62 & 0.00 & 0.38 \\ 0.32 & 0.00 & 0.68 \\ 0.08 & 0.00 & 0.92 \end{bmatrix} = \begin{bmatrix} 282 & 0 & 218 \end{bmatrix}$$

$$D_{2r*} = \begin{bmatrix} 282 & 0 & 218 \end{bmatrix} \begin{bmatrix} 0.83 & 0.00 & 0.17 \\ 0.00 & 0.00 & 0.00 \\ 0.36 & 0.00 & 0.64 \end{bmatrix} = \begin{bmatrix} 312 & 0 & 188 \end{bmatrix}$$

$$D_{3r*} = \begin{bmatrix} 312 & 0 & 188 \end{bmatrix} \begin{bmatrix} 0.55 & 0.00 & 0.45 \\ 0.00 & 0.00 & 0.00 \\ 0.12 & 0.00 & 0.88 \end{bmatrix} = \begin{bmatrix} 193 & 0 & 307 \end{bmatrix}$$

From Equation 5.21, it is apparent that there is considerable movement between the Above MC and Below MC performance categories from one iteration to the next. Movements from the Below MC performance category into the Above MC during all iterations indicates that this series of SOI is at least partially effective, concerning Transfer Reliability, for certain students in the Below MC performance category. The retention of students in the Above MC performance category in the second iteration exceeds that produced during the first and third iterations as do the movements from the Below MC to the Above MC performance categories. What was different in the development and delivery of the second SOI deserves further attention. Overall though, the Transfer Reliability of this series of SOI leaves much to be desired.

During the early stages of quality control implementation, it is important to remember that for many SOIs, regardless of how carefully designed and implemented, the Transfer Reliability may be dramatically lower than expected but not determining the Transfer Reliability will not change this situation.

5.12 Concluding Remarks

In discussions concerning solutions to poor reliability, there are three levels of variation reducing controls that are available to the organization: Engineered Controls, Administrative Controls and Instructor Dependent Controls. As mentioned in the introduction, these controls can be viewed as a hierarchy where Engineered Controls are always superior to either Administrative and/or Instructor Dependent controls. Instructor Dependent controls are the weakest form of controls and are essentially instructors doing what they are told to do when management is not physically present. Administrative controls consist of reconfiguring the existing mix of resources in a way that is believed to be more effective. Both Instructor Dependent controls and Administrative control require someone to physically check for compliance. Taken to extremes, checking for compliance can become a large and resource consuming activity. Engineered controls are the strongest form of controls and consist of controlling the method and the sequence in which new information is delivered and controlling the way in which the products of instruction are measured and stored. Engineered controls often employ passive data collection techniques in which measurements are collected, stored and aggregated automatically. On-line training is a form of Engineered Controls applied to instruction.

When reliability is low, regardless of the form of the reliability model, there are several structured strategies for the design and development of new or modified SOIs having adequate reliability. All of these strategies use some type of cause and effect analysis to identify the actions necessary to improve reliability such as Failure Modes and Effects Analysis. Improvement efforts are often coupled with either Quasi Experimental Designs or Design of Experiments methodology for validation. A Quasi Experimental design uses indicators such as persistent upward shifts in performance, when the measurements are displayed as a time series, to gauge the success of an improvement effort. When using Quasi experimental techniques all conceivable rival hypotheses must be explained away before the results are considered valid and so truly definitive conclusions are difficult to obtain. Quasi-experimental designs are discussed in Chapter 6 which concerns the Monitoring functions that are integral to quality control programs in general. DOE methodology, can in fact provide more definite conclusions because known rival hypotheses are controlled through specific design techniques, especially randomization, a feature not present in Quasi-experimental design techniques. DOE methodology is beyond the scope of this text but quasi-experimental methods are often satisfactory.

6

Monitoring

6.1 Introduction

Up to this point, we will have estimated both the accuracy and objectivity of the measurements resulting from the measurement system for each SOI or collection of SOIs. We will have also determined the minimum correct level of performance and/or the distributional specifications for each SOI. At this phase in the quality control implementation process, the organization should have migrated, where ever possible, to more frequent and less expensive passive measurements as opposed to less frequent and more expensive active measurements. The monitoring function is facilitated by passive data collection techniques.

Monitoring charts and graphs form the basis of the first pass data analysis regarding both instructional and affective outcomes. The most common type of chart for this purpose is known as a runs chart. In a runs chart, the measurements are plotted as a time series and compared to various probability boundaries that are also plotted on the chart. At the individual SOI level, the key aggregate statistics are the proportion of students achieving or exceeding the minimum correct level of performance, achieving or exceeding the true knowledge level of performance and/or conforming to the distributional specifications for the SOI. As mentioned in Chapter 5, the extent to which knowledge gained during instruction degrades over time is also of intense interest.

As mentioned in Chapters 2 and 3, there are four combinations of measurement characteristics that form a classification system for measurements: passive versus active data collection and in-situation versus post-situation measurement times. The type of measurement scale the measurements arise from and which of the four combinations of measurement characteristics describe the measurement system often points to the type of monitoring chart which is most useful.

An example of passive In-situation measurements arising from a ratio or interval scale are real time correlates of attention such as respiration rate and the galvanic skin response. Active in-situation measurements are most often based around an observational measurement system such as a rating scale. Observational measurements can approach the level of detail needed for an interval scale but are more commonly between an interval scale and an ordinal scale. Examples of a passive post-situation measurements are computer scored writing samples or computer corrected examinations. Passive post-situation measurements often contain the level of detail required for an interval scale and sometimes even a ratio scale. Active post-situation measurements are currently by far the most common types of measurement and involve instructors scoring examinations and assignments. These measurements often appear to approach the level of detail required for an interval scale but because of objectivity issues are typically only ordinal in detail.

The monitoring function can be expensive due to infrastructure requirements, so why do we get involved in this activity? Beyond the obvious advantage of providing information for managing the organization’s processes, measurements can point the way to causes of inadequate process performance. The concept of a cause for a particular case of inadequate performance is the entry way into a broader area of study known as Causation Theory.

When the Monitoring function is used in a manner informed by and consistent with Causation Theory, the result can be sustained incremental improvements in an organization’s processes. Causation Theory is a much larger topic than we have space to explore fully but what follows is a brief discussion of both past and current thinking regarding Causation Theory.

6.2 Causation Theory

When using a runs chart, the measurements will vary around a central value in either a random way or in non-random pattern that is assignable to some known or unknown cause. Random variation arises from complex relationships between both known and unknown variables. The nature of these variables are hopelessly entangled and inseparable. Assignable cause variation is the result of an intentional or unintentional change to the system for which an identifiable cause is either known, may be inferred or lies just beyond our current level of understanding. Beyond establishing and maintaining the status quo, the goal of the monitoring function is to aid in the detection of and correction of detrimental assignable cause variation. Assignable cause variation is often advantageous and the cause of this type of variation is also of interest. The measurements collected on the organizations processes should be sensitive to both types of assignable cause variation.

The word cause is often interpreted as meaning that a single cause, in the form of a force or stimuli, has intervened upon the pattern of the measurements. A detectable change in the pattern is known as an effect. The effect is most often ascribed to an intentional manipulation, often called a treatment or treatment condition, of some sort. From the perspective of causation theory, this use of the concept of cause and effect is somewhat simplistic.

For example, an intentional manipulation may not have caused the effect at all. The cause could be the result of an unrelated event coincident to the manipulation. Alternatively, the cause may not have even been an intentional manipulation but some ongoing and persistent influence such as gradual and inexorable changes in students over time. In order for a cause to be identified two conditions must be satisfied. The first is called the necessary condition and this refers to a condition that must be present in order for the effect to occur. The second is called the sufficient condition. A sufficient condition refers to a condition that will always produce the effect. An influence must be both necessary and sufficient in order to qualify as a cause. For this reason, the pathway for establishing causation focuses upon identifying a necessary condition or combination of necessary conditions that are sufficient to bring about the effect in a way that is both repeatable and reproducible.

6.3 The Four Cannons of Causation Theory

Four methods for establishing causation were first described by J. S. Mill in the late 1800's. These by and large still form the foundation of Causation Theory. The first of the cannons is known as the method of agreement and consists of studying the elements in common for several instances of an event of interest. The method of agreement is operationalized by employing some form of similarity analysis. A sophisticated method of similarity analysis is known as Variance Components analysis. In this method, measurements are collected and arranged according to membership in a hierarchy of the variables. The total common variance of these measurements is decomposed into distinct non-overlapping components of variation that, when added up, equal the total common variation. For example, if there are three variables A, B and C and each of these variables has two or more levels, all levels of B are present under each level of A and all levels of C are present under each level of B under each level of A. In this way, the variables and levels of variables are nested in a factorially exhaustive hierarchy. Once the variance components are calculated, the output of the analysis is communicated as the percentage of the total common variation attributable to each component of variation. The percentage of the total common variation that remains unexplained equals the total common variation minus the total variation attributable to the variables and this is also presented as a percentage in the results. Once a particular combination of variables that exhibit excessive variation is identified specific techniques are used to tease out the cause. Techniques that are universally useful for similarity analysis are brainstorming, flowcharts, cause and effect diagrams such as fish bone diagrams, Pareto charts, histograms among others.

The second cannon is known as the method of difference. In the method of difference, the objective is to identify a causal relationship by observing the differential effects resulting from two situations that are alike in all respects except for the presence or absence of a treatment condition. The critical phrase here is ‘alike in all respects’. The extent to which this alikeness can be verified limits the number of alternative explanations for the effect that must be eliminated before interpreting the relationship between the cause and the effect. In general, alternative explanations must be either controlled by study design techniques or if not explicitly controlled, must be either accepted as possible causes or explained away in a truly plausible way. With regard to the efficacy of instruction, there are several alternative explanations lurking in the background to which all studies are susceptible and these are known as extraneous variables. The most important extraneous variables in instructional studies are as follows:

History: These are events other than treatment conditions, that occur between the beginning and the end of the study. One strategy used to control history effects is to keep studies relatively short.

Maturation: These are conditions within the individual student that change over time. A strategy for controlling maturation effects is to avoid long studies that last several months or years. A Repeated Measures study design offers another way of controlling maturation effects on the measurements from both short and long duration studies involving the same set of students.

Mortality: This refers to the loss of one or more individuals from a study. To counter this, increase the sample size to accommodate for the expected number of individuals that may not complete the study.

Instrumentation: This refers to changes in the ability to measure the response. These can include the discontinuation of instructional software or of a standardized test critical to the study. Instrumentation effects can also result from changes in an instructor generated examination used in the study. Also, any changes that affect the accuracy and objectivity characteristics of the measurement system can result in measurement effects. Planning for the stability of the measurement systems used in a study is the best control method for instrumentation effects.

Selection: A selection effect results from bias in the assignment of students to treatment conditions. The best counter strategy for selection effects is random selection of students and random assignment of students to treatments. Another is to design studies in ways that allow students to be their own controls using a Repeated Measures study design.

Sequencing: These are changes attributed to participation of the student in more than one treatment condition. Many times determining the best sequence of treatments in which to present information is often a goal of a study. When this is not the case, random selection and assignment to treatment conditions is the best counter strategy. A design technique known as counterbalancing, in which half of the students receive treatments in the opposite order is also effective in identifying sequencing effects. For example, half of students receive the treatments in ABBA order and the other half receive the treatments in BAAB order.

Sophistication: These are changes attributed to familiarity with the study procedures. Familiarity with study procedures can be a goal of a study. However, in cases were sophistication is not a goal, making sure that instructions are understood is critical. Also, verifying that antecedent skills and knowledge are present can control sophistication effects. Students developing test savvy methods such as using partial knowledge on multiple choice questions is a form of sophistication effect.

Statistical Regression: This refers to the tendency of individuals selected on the basis of membership in an extreme group to return to a central level of performance. One counter strategy is to avoid selecting students based upon membership in extreme groups.

Participant effects: These are effects attributed to changes in motives, motivation or attitudes of participants. It is best to shield participants from details of the study. Studies that involve excessive repetition of tasks are most subject to participant effects.

Principle effects: These refer to changes in participants that result from principle actions. Examples include differences in the way in which the principle interacts, answers questions or provides instructions that are inconsistent from one treatment group to another. Random selection of instructors, focusing on variables with large effect sizes, making planned studies part of routine operation and using automation are all proven counter strategies.

The third cannon is known as the joint method of agreement and difference. In this pathway, elements in common are first explored using one or more methods of agreement. From the elements in common, one or more hypotheses are developed that are then tested using a method of difference. The joint method of agreement and difference provides a proven pathway for establishing causation when the treatment group and control group are matched or are shown to be equivalent. A common situation encountered during improvement efforts occurs when an alternative version of an SOI is the treatment condition and the existing SOI is the control. In this case, both the treatment and control SOIs are shown to have all important variables in common. Then one group is randomly assigned to the treatment condition.

The fourth cannon is know as the method of concomitant variation. This cannon may be viewed as simply an extension of the joint method of agreement and difference that can accommodate several treatment conditions. In this method, once one or more methods of agreement have been used to formulate testable hypotheses, rather than exploring the simple presence or absence of a single hypothesized cause, the effects of several different amounts of treatment conditions, also known as factor levels, are tested simultaneously. The advantage of the fourth cannon is that unique combinations of perspective necessary conditions can be studied systematically and exhaustively in a single, albeit, more complicated study. The fourth cannon is the only one that allows combinations of necessary conditions to be studied in hopes of establishing a combination of necessary conditions that together form a sufficient condition.

6.4 The Concept of a Root Cause

The concept of a cause is subordinate to the concept of a root cause. In the face of incomplete information, the most common circumstance encountered outside of a laboratory environment, to state that the root cause has been identified, a complete explanation of the occurrence of the effect must be obtained in which no change in the explanation will ever occur. Due to the high standard of proof required for identification of a root cause, the predominant line of thinking regarding causation is that the concept of a cause is so wedded to the level of understanding existing at the time, that a hypothetical root cause is never a complete explanation but only a hypothesis that has withstood all challenges to date. This concept is known as the Position of Falsification. The position of falsification is the central concept of statistical hypothesis testing. In this way a hypothesized cause is never confirmed by a study but rather several studies only fail to disconfirm the hypothesis. After many successful challenges, a particular hypothesis may come to be relied upon as factual.

From the standpoint of improving the efficacy of instruction, the position of falsification implies that a cause is actually only the isolation of a single cause and effect relationship embedded in a larger network of cause and effect relationships. For this reason, any study on the efficacy of instruction must go to extreme lengths to hold the larger network of cause and effect relationships constant while studying the relationship or relationships of interest. The monitoring function is the most common provider of evidence concerning the stability of the causal network. The monitoring function also can provide evidence in support of an effect resulting from a causal improvement effort.

As mentioned earlier, the most common type of charts used in the monitoring function of a quality control program are known as runs charts. There are several variations on the runs chart theme but all plot the passage of time along the x-axis and the response on the y-axis. There are charts for both individual measurements and for collections of measurements and these will be explained separately when necessary.

6.5 Charts for Individual Measurements

In the most conservative case, the probability that a question will be answered correctly is viewed as a binomial experiment and the binomial distribution becomes the appropriate model. Of the statistical control chart types, the one designed for measurements communicated as proportions, for use when the number of questions on each examination is approximately equal, is the p Chart. The p Chart is most applicable to examination scores when the proportion of correct answers is large, on the order of $p = 0.95$. This is because the standard deviation will be relatively small. The p Chart is also most appropriate when the number of measurements available for use in calculating the control limits is approximately 30 or more. The standard deviation of a binomial distribution is completely determined by the average proportion of questions answered correctly p. This is because the variance of the binomial distribution equals $p \times q$ where $q = 1 - p$. Equation 6.1 is used to estimate the standard deviation of the measurements for the p Chart. The standard deviation is in turn used to calculate the +/- 1, 2 and 3 sigma control limits. The measurements are then plotted as a time series between the control limits.

Equation 6.1: Standard Deviation of a Proportion

$$\sigma_{\bar{p}} = \sqrt{\frac{\bar{p} \times (1 - \bar{p})}{\bar{n}}}$$

For example, if the average proportion of questions answered correctly equals p = 0.8, the average proportion of incorrect answers q equals 1 - p = 0.2. If the average number of questions on each examination equals 25, the standard deviation equals $((0.2 \times 0.8) / 25)^{0.5} = 0.08$. The +/- 1, 2 and 3 sigma control limits for the p Chart for these measurements equals 1.04, 0.96, 0.88, 0.72, 0.64 and 0.56. Notice that the control limits are all reasonable with the exception of the + 3 sigma control limit which equals 1.04. This control limit is adjusted downward to 1.00. Unfortunately, the 2 and 3 sigma lower control limits for the p Chart are often ineffectually low and are most commonly below the minimum correct level of performance.

Treating an individual examination score as a binomial experiment is the strictest interpretation of these types of measurements. Examination scores are the result of volition on the part of test takers and thereby contain information beyond the sum of a blindly random dichotomus process. For this reason, the X and Moving Range Chart, abbreviated X MR, should be considered as a method of monitoring individual measurements. The advantage of using the X MR chart is that the control limits and especially the lower control limits are typically more useful than those resulting from the p Chart. Another reason for selecting the X MR chart is that the MR portion of the X MR control chart is especially sensitive to shifts in the measurements.

The upper and lower control limits for the X portion of the X MR control chart are calculated from the average of the measurements plus a constant multiplied by the average moving range of these measurements. The results from multiplying the average moving range by the correct constant E_2 equals the +/- 3 sigma control limits for the measurement average. The + 3 sigma control limit for the range average is estimated by multiplying the moving range average by the correct constant D_4.

For situations in which the moving range is calculated from n = 2 measurements, Equations 6.2a and 6.2b are used to calculate the upper and lower control limits, respectively. The constant used when the moving range is calculated from n = 2 measurements for the X portion of the X MR chart is E_2 = 2.659. The constant for the moving range portion of the X MR chart when n = 2 measurements are used to calculate the moving range is D_4 = 3.267. Constants are available for n = 2 to 10 measurements in the moving range. The control limits resulting from the use of these constants are 3 standard deviations from the average of X or from 0 for the moving range. In order to obtain an estimate of the standard deviation for the X chart portion of the X MR chart, the distance from the average to a control limit is divided by 3. For example, the +1 and +2 sigma upper control limits for the X portion of the chart are placed UCL / 3 and 2 x (UCL / 3) above the average of X, respectively. The -1 and -2 sigma lower control limits are placed at UCL / 3 and 2 x (UCL / 3) below the average of X, respectively. The 1 and 2 sigma control limits for the moving range portion of the X MR control chart are calculated in a similar way using the 3 sigma upper control limit calculated using Equation 6.2c. The 1 and 2 sigma control limits for the MR portion of the chart are placed at 0 + UCL_{MR} / 3 and 0 + 2 x (UCL_{MR} / 3) ,respectively.

Equation 6.2a: Upper Control Limit for the Average of X for n = 2, X MR Chart

$$UCL_x = \bar{x} + E_2 \times \overline{MR} = \bar{x} + 2.659 \times \overline{MR}$$

Equation 6.2b: Lower Control Limit for the Average of X for n = 2, X MR Chart

$$LCL_x = \bar{x} - E_2 \times \overline{MR} = \bar{x} - 2.659 \times \overline{MR}$$

Equations 6.2c: Upper Control Limit for the Average MR for n = 2, X MR Chart

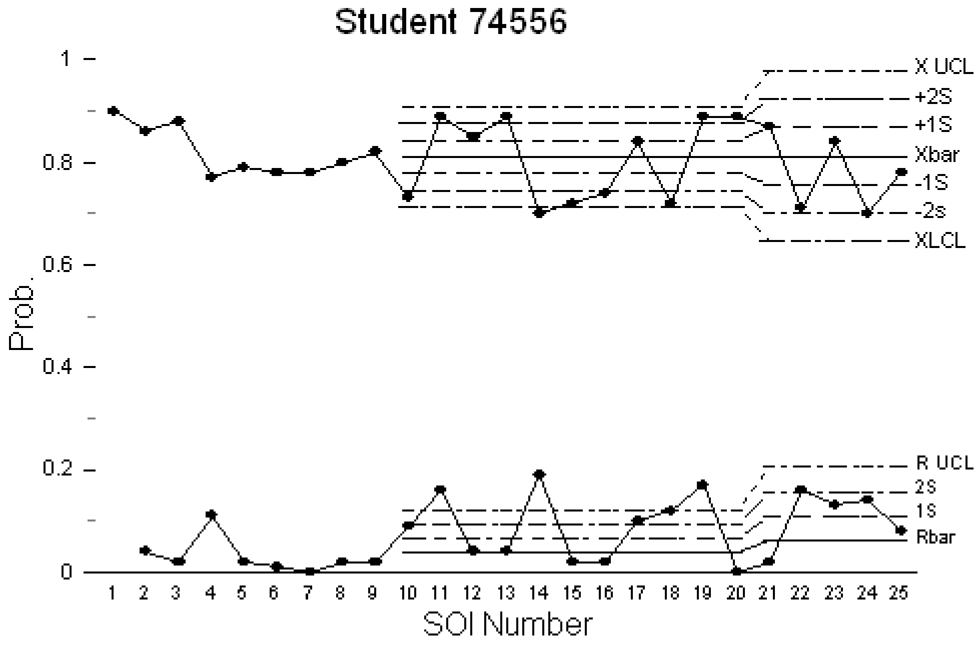
$$UCL_{MR} = \overline{MR} + D_4 \times \overline{MR} = \overline{MR} + 3.267 \times \overline{MR}$$

The X MR control chart requires a minimum of 20 measurements in order to estimate the control limits. This restricts the use of this chart type to situations where a large number of measurements are available. To counter the requirement for 20 measurements, some authors have suggested that as few as 10 measurements are adequate as long as control limits are recalculated frequently or until stable estimates are obtained. One problem with this approach is that the Shewhart constants, such as E_2 and D_4 will produce control limits that are much too narrow when only a small number of measurements are available for calculating the control limits. This is because the variation of the population has not been captured by the range of the available measurements. To illustrate this point refer to Figure 6.1 which contains an X MR chart for the 25 measurements contained in Table 6.1.

Table 6.1: The 25 Measurements for Figure 6.1

SOI	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Score	0.90	0.86	0.88	0.77	0.79	0.78	0.78	0.80	0.82	0.73	0.89	0.85	0.89
SOI	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	
Score	0.70	0.72	0.74	0.84	0.72	0.89	0.89	0.87	0.71	0.84	0.70	0.78	

Figure 6.1: X MR Chart for Measurements from Table 6.1(control limits calculated at n = 10 and n = 20 measurement)



The control limits for Figure 6.1 were calculated using the first 10 measurements and then recalculated using the first 20 measurements. The average moving range for the first 10 measurements equals 0.037 and for the first 20 measurements equals 0.063. This is because, the first 10 measurements are uncharacteristically stable when compared to the next 10 measurements. In this figure, the control limits calculated from the first 10 measurements are inappropriately narrow and will lead to incorrect interpretation of measurements 11 thru 20.

6.6 Patterns Indicating Assignable Cause Variation

Once satisfactory control limits are in place, the next step is to understand the runs chart in light of assignable cause variation v.s. random variation. There are several patterns indicating the presence of assignable cause variation collectively known as the WECO rules and these are listed below. Each pattern description has a definition of the pattern, the name of each pattern in parentheses followed by the maximum probability of occurrence of the pattern.

1 point beyond a 3 sigma control limit (Out of Control) p = 0.00135

2 out of 3 successive measurements on the same side of the mean line beyond the 2 sigma limit. (Freak) p = 0.0016

4 out of 5 successive measurements on the same side of the mean line beyond the 1 sigma limit. (Freak) p = 0.0028

8 successive point on the same side of the mean line. (Shift) p = 0.0039

7 successive measurements either collectively rising or falling (Run) p < 0.0078

10 of 11successive measurements either collectively rising or falling, (Trend) p < 0.0054

14 successive measurements lying within +/- 1 sigma of the mean line. (Stratification) p = 0.0048

For example, the probability for the pattern 4 out of 5 successive measurements landing beyond the 1 sigma control limit on the same side of the average line is calculated as follows: There are 5 ways in which this pattern can occur: 11110, 11101, 11011, 10111 and 01111. The 1's represents sequential measurements landing beyond the 1 sigma line on either one side of the average line or the other but not both. A 0 represents a measurement landing within the 1 sigma line or on the other side of the center line from the all 1's in the pattern. In the normal distribution, 16% of the area under the distribution is contained in each tail of the distribution beyond 1 sigma. The probability of the first of the five ways in which the 4 out of 5 pattern can occur equals 0.16 x 0.16 x 0.16 x 0.16 x 0.84 = 0.16^4 x 0.84 = 0.00065536 x 0.84 = 0.0005505. There are a total of five ways in which this pattern can occur so the total probability equals 5 x 0.0005505 = 0.00275 ~ 0.0028.

Other patterns indicating assignable cause variation are also possible. There can also be recurring cycles which produce a sinusoidal pattern and stable mixtures in which successive measurements vary abruptly about the center line. Abrupt variation about the center line occurring in an alternating or somewhat balanced way indicates that two or more distributions are being plotted on the same chart. A random pattern of extreme departures from the average line indicates the presence of extreme Freak measurements, sometimes called fliers.

6.7 The Weighted X Control Chart

For instances when too few successive measurements are available or even possible from which to calculate control limits, such as is the case for a short series of SOI’s, a statistical control chart known as a Weighted X abbreviated wtX Chart is available. The control limits for the wtX control chart are calculated using Equation 6.3. You will notice that Equation 6.3 is actually a modified version of Equation 6.1 in which the TK level of performance is the target p and n is the cumulative number of questions.

The benefit of this control chart type is that, even though the number of questions on a typical examination is too small to result in meaningful control limits, the cumulative number of questions becomes large enough for meaningful control limits to result. Using the cumulative number of questions in the divisor results in control limits that become narrower for each successive weighted average score plotted on the chart. Other relevant information can be plotted on the wtX Chart such as the TK level of performance, the individual examination scores and the MC level of performance for each SOI. Example 6.1 illustrates the details of the wtX Control Chart.

Equation 6.3: Sigma for the wtX Chart

$${}_{wt}\sigma_{TK} = \sqrt{\frac{{}_{wt}TK_{pr.}(1-{}_{wt}TK_{pr.})}{\sum n}}$$

Example 6.1

Once the number of questions on each examination in a series of SOI's are known, construction of the wtX chart begins by calculating the weights based upon the total number of questions contained in all examinations used in the calculation of a particular weighted value. The weight is applied to the scores and all weighted control limit values used in the construction of the chart. For example, the weights for the third weighted average score and all weighted control limits and other weighted values of interest for this weighted average are calculated using the number of questions on the first three examinations. The number of questions on the first three examination equal 23, 31 and 29 respectively. The weights for the first three examination scores equal 23 / (23 + 31 + 29) = 23 / 83 = 0.2771, 31 / 83 = 0.3735 and 29 / 83 = 0.3494, respectively. Using weights is a bit tedious but these calculations remain the same for a given collection of SOI's as long as the number of questions on each examination does not change. If one or more examination sizes change, the table of weights will need to be recalculated. Attention to details is especially important anytime we ask personnel to modify their methods in order to improve an SOI based upon feedback from a control chart.

In this example, a chart is prepared for a series comprised of 10 SOI's having the MC levels of performance and number of questions shown in Table 6.2. Calculation of the table of weights proceeds by listing the number of questions in columns as shown in Table 6.3. Dividing the number of questions on each examination by the cumulative number of the questions up to that point results in the table of weighs shown in Table 6.4.

Table 6.2: Examination Information

SOI	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
wtTK	0.850	0.839	0.843	0.851	0.842	0.842	0.838	0.839	0.843	0.840
MC	19	25	23	23	31	20	25	25	23	34
n	23	31	29	28	40	25	32	32	28	42
Score	17	26	28	20	25	19	30	29	26	40

Table 6.3: Cumulative Number of Questions

					SOI					
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
# of Ques.	23	23	23	23	23	23	23	23	23	23
	0	31	31	31	31	31	31	31	31	31
	0	0	29	29	29	29	29	29	29	29
	0	0	0	28	28	28	28	28	28	28
	0	0	0	0	40	40	40	40	40	40
	0	0	0	0	0	25	25	25	25	25
	0	0	0	0	0	0	32	32	32	32
	0	0	0	0	0	0	0	32	32	32
	0	0	0	0	0	0	0	0	28	28
	0	0	0	0	0	0	0	0	0	42
Cum # Ques.	23	54	83	111	151	176	208	240	268	310

Table 6.4: Table of Weights

				SOI					
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1.000	0.426	0.277	0.207	0.152	0.131	0.111	0.096	0.086	0.074
0.000	0.574	0.373	0.279	0.205	0.176	0.149	0.129	0.116	0.100
0.000	0.000	0.349	0.261	0.192	0.165	0.139	0.121	0.108	0.094
0.000	0.000	0.000	0.252	0.185	0.159	0.135	0.117	0.104	0.090
0.000	0.000	0.000	0.000	0.265	0.227	0.192	0.167	0.149	0.129
0.000	0.000	0.000	0.000	0.000	0.142	0.120	0.104	0.093	0.081
0.000	0.000	0.000	0.000	0.000	0.000	0.154	0.133	0.119	0.103
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.133	0.119	0.103
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.104	0.090
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.135

When the student’s scores, contained in Table 6.5, are multiplied by the weights from Table 6.4 and the columns are totaled, the results are the weighted average scores as shown in last row of Table 6.6. The same procedure is used to calculate the weighted average TK level of performance or any other weighted time series to be included on the wtX chart.

Table 6.5: Scores for an Individual Student

SOI	1	2	3	4	5	6	7	8	9	10
Score	17	26	28	20	25	19	30	29	26	40
Score pct.	0.74	0.84	0.97	0.71	0.63	0.76	0.94	0.91	0.93	0.95

Table 6.6: Weights x Scores

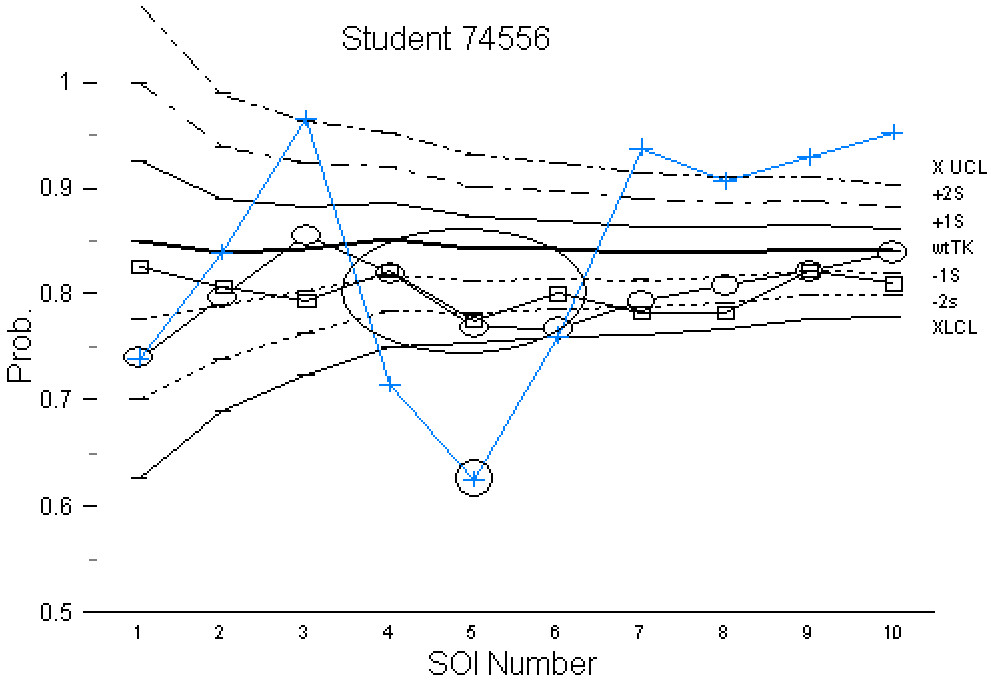
					SOI					
	1	2	3	4	5	6	7	8	9	10
0.74	0.74	0.31	0.20	0.15	0.11	0.10	0.08	0.07	0.06	0.05
0.84	0.00	0.48	0.31	0.23	0.17	0.15	0.13	0.11	0.10	0.08
0.97	0.00	0.00	0.34	0.25	0.19	0.16	0.13	0.12	0.10	0.09
0.71	0.00	0.00	0.00	0.18	0.13	0.11	0.10	0.08	0.07	0.06
0.63	0.00	0.00	0.00	0.00	0.17	0.14	0.12	0.10	0.09	0.08
0.76	0.00	0.00	0.00	0.00	0.00	0.11	0.09	0.08	0.07	0.06
0.94	0.00	0.00	0.00	0.00	0.00	0.00	0.14	0.13	0.11	0.10
0.91	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.11	0.09
0.93	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.08
0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13
wt.	**0.74**	**0.80**	**0.86**	**0.82**	**0.77**	**0.77**	**0.79**	**0.81**	**0.82**	**0.84**

If you'll recall, the null assumption of the True Knowledge model is that each student is a member of the True Knowledge distribution. So, on the wtX chart, the control limits are based upon the weighted TK level of performance rather than the weighted average student scores. Using the TK level of performance eliminates the need for a baseline. Control limits and the associated assignable cause variation patterns provide evidence either in support of or against the student being a member of the True Knowledge distribution. Using Equation 6.3 to calculate the 1, 2 and 3 sigma control limits results in the control chart shown in Figure 6.2.

Notice that the unweighted scores for this student drop dramatically at the 4th SOI in the series. This pattern of poor performance continues until the 7th SOI. The location of these scores results in an assignable cause variation pattern, also called an out of control condition, for the weighted average scores. These points are enclosed in a large ellipse. The out of control condition is a Freak pattern in which 2 of 3 points on the same side of the centerline are beyond the 2 sigma control limit. This pattern indicates non-random variation that is inconsistent with the student being a member of the True Knowledge distribution. If the student is being reassessed and receiving remedial instruction after each SOI in which the score is below the MC level of performance, these additional actions are clearly ineffective. The out of control condition exhibited by the weighted average score should result in a strenuous intervention.

The weighted average score for the 10th SOI in the series is equal to the weighted average TK level of performance. Due to the fact that the weighted average student score for the 10th SOI is above the weighted average TK level of performance, the student has in some sense meet the minimum required level of performance for the series of SOI's. However, because of the amount of intervention required, improvement efforts should be considered if this is a pattern exhibited by several students.

Figure 6.2: Weighted Average Chart



Legend for Figure 6.2

Bold Center Line: This line represents the weighted TK level of performance for each SOI in the series.
Line with Ellipses: This line and pattern represents the weighted scores for student 74556.
Line with Squares: This line and pattern represents the weighted MC level of performance.
Line with Cross Hatch: This line contains the unweighted scores signal for the student.

This concludes our discussion of control charts that are appropriate for individual measurements. There are many more types of charts and ways of displaying data that are useful for specific purposes. Many time ad-hoc charts are developed in order to focus on specific goals. Certain additional runs chart types and techniques are presented in the following discussion of specific types of planned studies.

6.8 Planned Studies

When out of control conditions persist, there are basically two approaches for bringing a process into control. The first approach involves raising the mean or median through direct actions on the part of the persons developing and implementing instruction. These activities typically include the application of instructor dependent controls and/or administrative controls. The second approach is to reduce variation of the process while simultaneously raising the mean or median of a time series. Variation control is much more difficult and typically requires both additional resources and critical innovations. For this reason, a balanced strategy is to determine the level of improvement required using the appropriate reliability model from Chapter 5 and improve the one or more SOIs offering the greatest level of improvement in centering and variation.

The results from reliability modeling should be proposals for improvement efforts. Of these, one or more are selected for study and this leads to the formation of one or more testable hypotheses. The position of falsification is commonly implemented in instructional applications through the use of one or more runs charts. The runs charts containing measurements of the response are typically either Single Subject Studies, Quasi-experimental designs or applications of Design of Studies (DOE) methodology. The main difference between DOE methodology and quasi-experimental methods is that random selection and random assignment of students to treatment and non-treatment groups is both possible and accomplished in valid applications of DOE methodology. This difference is not trivial because randomization is the best defense against the sources of extraneous variation discussed earlier.

In practice, the most common method for forming groups of students is thru non-random assignment based upon who is available when an SOI or collection of SOIs is offered. The discussion that follows concerns how to use runs charts in ways that meet the requirements of single subject designs and quasi-experimental designs. When design concepts are married to runs charts, every point on a particular runs chart becomes an application of the position of falsification. Conformation of the efficacy of an improvement effort is demonstrated by a predicted and sustained shift in the measurements contained in one or more relevant runs charts.

6.9 Single-Subject Designs

When measurements of an individual student’s progress through a series or category of SOIs is tracked over time in order to determine the efficacy of an improvement effort, this becomes a single subject design. Single subject designs form the foundation upon which instructional quality control programs are built. In this type of design, there are two patterns of performance that are of most interest. The first and most compelling pattern is a predicted and sustained step function improvement in a relevant time series. When a treatment condition is introduced at a known point in the time series, the design is known as an Interrupted Time Series design. Interrupted time series designs are described using the notation B-B-A-A where, B’s are the baseline measurements and A’s represent measurements occurring after Treatment A.

In some studies, it is possible to apply a treatment and then withdraw the treatment. In certain circumstances, the application and withdrawal sequence can be repeated several times. The pattern for this type of study is a step function improvement from the baseline followed by a return to the baseline once the treatment condition is removed forming an B-A-B patten. If the treatment is re-instated the predicted pattern becomes B-A-B-A.

Establishing the baseline level of performance is somewhat subjective but an agreed upon definition is a continuous set of measurements that exhibit no trend and contain a relatively low level of variability. A further criteria could be that no violations of the WECO rules are present during the baseline period. The need to establish a baseline limits the usefulness of single subject designs for short series or categories of SOI's if examination scores are the measurements. This is because too few measurements are available to establish a stable baseline level of performance .

There are no definitive methods for establishing a baseline for short series of SOI's but there are a few methods for approximating the baseline level of performance with only a limited number of measurements. One method involves approximating the maximum standard deviation using the range of the available measurements. Another, which requires a few more total measurements, uses the distance between the 25th and 75th percentile locations to identify outliers which may in turn be treatment effects. The method selected for establishing a baseline depends upon the number of measurements available where more measurements provide more confidence in the baseline.

Estimating the maximum standard deviation requires 3 to 6 measurements and is accomplished using Equation 6.4. The maximum small sample standard deviation is then used to estimate one or more control limits. A word of caution for use of Equation 6.4 is that the estimate of the standard deviation can become an overestimate after about 8 measurements.

Equation 6.4: Maximum Standard Deviation for Small Sample

$$S \leq \frac{R}{2}\sqrt{\frac{n}{n-1}}$$

For example, using Equation 6.4 for the baseline scores 0.83, 0.76 and 0.73, the estimated maximum standard deviation equals 0.061. For these baseline scores, the 2 sigma control limits equal 0.773 ± 2 x 0.061 or 0.651 to 0.895 and the 3 sigma limits are 0.773 ± 3 x 0.061 or 0.59 to 0.956. A tentative positive treatment effect occurs if the first measurement subsequent to the treatment is above 0.956 or if the next 3 points are above the center line and 2 of 3 are above 0.895.

Equation 6.5: Standard Deviation for 0.83, 0.76 and 0.73

$$S \leq \frac{0.10}{2}\sqrt{\frac{3}{3-1}} = 0.061$$

When 9 to 15 measurements are available it is possible to estimate the inter-quartile range, IQR, of the measurements. For baseline samples of 15 measurements or less, interpolation is used for estimating the 1st and 3rd quartile locations. The IQR is then used to either calculate the top and bottom hinge locations for a Box Plot. The top hinge is located at the first measurement less than 1.5 IQR above the third quartile and the bottom hinge is located at the first measurement larger than 1.5 IQR below the first quartile. A tentative positive treatment effect occurs when the measurements subsequent to the treatment are consistently above the top hinge.

For example, if in addition to the three scores of 0.83, 0.76 and 0.73, the next six scores are 0.83, 0.79, 0.79, 0.76, 0.62 and 0.76, the median equals the middle measurement when the scores are sorted in either ascending or descending order. The median, in this case equals 0.76. A single measurement will exist at the median for samples containing an odd number of measurements. For even numbers of scores, the central two measurements are averaged to obtain an estimate of the median.

The nine scores shown in descending order are: 0.83, 0.83, 0.79, 0.79, 0.76, 0.76, 0.76, 0.73 and 0.62. Both the 25th and 75th percentile locations are arrived at by interpolation and equal (0.76 + 73) / 2 = 0.745 and (0.83 + 0.79) / 2 = 0.81, respectively. The interpolated IQR equals 0.81 - 0.745 = 0.065. The top hinge equals the first value lower than 0.81 + 1.5 x 0.065 = 0.81 + 0.0975 = 0.9075 which equals 0.83. The bottom hinge equals the first value larger than 0.745 - 1.5 x 0.065 = 0.745 - 0.0975 = 0.6475 which equals 0.73.

Setting hinge locations at then next value closer to the median than 1.5 x IQR is meant to allow asymmetry in the location of IQR box about the median to affect the hinge location. This allows Box Plots to be robust even for skewed distributions.

However, for small samples locating the hinges at the nearest value closer to the median results in hinges that are inappropriately narrow. This can lead to a misinterpretation of the data leading to a conclusion that a treatment effect occurred when in fact a treatment effect is not present. For this reason, interpolating the distance between +/- 1.5 x IQR and the next value closer to the median results in more meaningful hinge locations. Using this approach, the upper hinge location for the 9 measurements equals (83 + 0.9075) / 2 = 0.869 and the lower hinge equals (73 + 0.6475) / 2 = 0.689.

According to this analysis, the score 0.62 is an outlier that must be explained before the baseline is considered stable. In this application of the Box Plot, measurements subsequent to the treatment that are persistently beyond one hinge or the other but not both are exhibiting a non-random pattern that supports the existence of a treatment effect.

Now that we have discussed the concepts necessary for establishing a small sample baseline level of performance for the interrupted time series pattern as well as how to demonstrate a tentative treatment effect when only a small number of measurements are available, our discussion turns to quasi-experimental designs appropriate for aggregations of students.

6.10 Quasi-experimental Designs

The most common way in which groups of students are currently aggregated in instructional settings is by accommodating all or most of the students available when an SOI or collection of SOIs is offered. This fact typically precludes the ability to randomly assign students to treatments. This is unfortunate because this rules out the use of many DOE methods which allow more complicated combinations of variables to be included in a planned study. However, the method of difference can still be applied if all outside influences can be either controlled or at least held constant during the study.

There are three designs that are of interest for our purposes: the Interrupted Time Series Design, the Non-Equivalent Control Group Design and the Cohort Design. Of the three, the Interrupted Time Series Design is the simplest because the same group of students are monitored over time. This allows the group to act as it’s own control group. When the group acts as it’s own control, Repeated Measures non-parametric statistical tests, such an the Friedman test and the Van der Waerden Normal Scores test, can be used to analyze the measurements. The Non-equivalent Control group design compares the treatment group to a matched control group. In this design, even though the groups are matched, the groups are still considered Non-equivalent. The Cohort design is used when the group of students is completely new at each measurement of the performance of an SOI. This is the most common situation for groups of students completing independent SOIs.

6.11 The Interrupted Time Series Design

The Interrupted Time Series design requires that a baseline be established and a treatment effect is evidenced by a sustained shift in the time series after introduction of the treatment condition. In instructional applications, the number of students in the group being monitored changes slightly between measurements as some students are absent, move or withdraw from the group. What we need is a method that uses all of the measurements for comparing the groups performance over time in order to determine if improvement efforts are effective.

In this design, the group acts as it's own control and so most sources of extraneous variation are controlled. This control mechanism becomes more tenable when the baseline is stable before the improvement effort.

Most statistical control charts suitable for groups of students require that the number of students represented by each point on the control chart be constant. One strategy for accomplishing constant sample size is to randomly select a subgroup of students containing the same number of students from each group of students. The average for each subgroup is used to develop an Xbar and Range control chart. In this way, the WECO rules can be used to detect assignable cause variation that may support a treatment effect. This strategy is inferior because all of the measurements available are not used. Any time the entire group is measured at each occasion, runs charts comprised of various percentile locations have proven the most useful. This is because we have measurements on the entire population and so using parameters, such as the standard deviation, to describe the population is not necessary.

This first set of percentile locations seeks to heavily scrutinize the edges of the score distribution. Common examples are 0.01, 0.025, 0.05, 0.10 and 0.20 and 1 minus these the locations or 0.80, 0.90, 0.95, 0.975 and 0.99. These percentile locations are useful when comparing the group against a criterion such as the MC level of performance. By displaying the MC proportion of questions on the run chart, the proportion of students requiring reassessment and remedial instruction can be estimated for each SOI included on the runs chart. Also, the shape of the distribution can be evaluated for each SOI which provides information concerning conformance to distributional specifications. An interesting point is that locations 0.20 and 0.80 are very robust indicators of spread. The stability of these two locations, known as the inter-decile 80%, abbreviated ID80, even extends to rare instances in which the population distribution is U-shaped.

Another common set of percentile locations, and one that works well for larger aggregations of students, corresponds to the median +/- the percentile equivalents of 1, 2, and 3 sigma probability boundaries. In this runs chart theme, the probabilities appearing from bottom to top on the runs chart are 0.0013, 0.023, 0.341 and 0.659, 0.977 and 0.9987. The advantage of using these probabilities is that, in a general sense, these probabilities have the same interpretation as the 1, 2 and 3 sigma control limits used on conventional statistical control charts.

When comparing several instances of instruction, where the group size is 9 students or more, Box Plots provide an excellent visual representation of the distributional characteristics of the measurements. A variation on the Box Plot theme is to define the box using the two sided 95% confidence interval for the median rather than the IQR. An example of this chart type is contained in the discussion of the Nonequivalent Control Group design.

6.12 Nonequivalent Control Group Design

In the Non-equivalent Control Group design, both the control group and the treatment group have a known quantity of the variable or variables of interest before the administration of the treatment condition. Establishing the quantity of the variable or variables of interest is accomplished either through a pretest, in the case of an independent SOI, or by one or more prior measurements in the case of a series of dependent SOI's. The prior measurements need not be limited to examination scores but can include any measurements helpful in matching students. The study involves exposing the treatment group to the modified conditions and the control group to the status-quo. A classic example of the non-equivalent control group design is illustrated Figure 6.3. In this figure, Instructor a is an existing successful instructor and Instructor b is a new instructor being evaluated for bench marking efforts. The expectation is that the new instructor's performance will be equal to that of the experienced instructor.

The Non-equivalent Control Group design is susceptible to local history, statistical regression and selection/maturation influences. A local history effect could be as simple as the students preferring Instructor a and realizing that the two instructors are being compared using a study. A local history effect occurs If the students collude to manipulate the outcome of the study.

Statistical regression toward an as yet unknown group mean is problematic if the techniques used to identity the treatment or control groups select students from even moderately divergent groups. In fact, it has been shown that even when great care is taken in selecting the treatment and control groups, through matching in every possible respect, the control group and the treatment groups very often diverge even when the instruction received for both groups is identical. For these reasons, the most compelling patten indicating a treatment effect, when statistical regression is being controlled, occurs when the group that scores the lowest on the pretest, baseline or criterion is selected as the treatment group and this groups performance exceeds that of the control group after the treatment. For example, if the treatment condition is an improved SOI, that is expected to be superior to the status quo, the group with the lower pre-test, baseline or criterion score is assigned to the improved SOI. If performance of the treatment group surpasses that of the control group, lines connecting the pre and post scores for the two groups will cross. This is know as a cross-over effect. In the cross-over effect, statistical regression operating as a rival hypothesis can be explained away because it is difficult to imagine a situation where the treatment group would regress enough to become significantly higher than the control group.

In a selection/maturation effect students initially having more of a variable obtain more of the variable at a faster rate than those starting with a lesser quantity of the variable. The cross-over patten is the exact opposite of a selection/maturation effect because the group initially having less of the variable must gain more of the variable, at a faster rate, than the group initially having more of the variable in order to cross-over. If possible, it is best to assign the treatment to the group having the pre-test or criterion score for which a crossover effect rules out selection/maturation or statistical regression effects.

There are several other possible patterns that can arise in the Non-equivalent Control Group design for situations in which more of the treatment variable is desirable. The first is the Increasing Treatment Effect I pattern. This pattern occurs when the treatment group has a pretest, baseline or criterion score above the control group and the treatment group increases while the control remains stable. This is a study design flaw. However, assigning the treatment to the group having the higher pre-treatment score is sometimes desirable. An example of this decreasing treatment effect will be shown in Example 6.2. If the increasing treatment effect I pattern will not strengthen the conclusions, the potential for this pattern should be avoided by assigning the treatment to the group with the lower pre-test, baseline or criterion score. If this pattern is encountered, both selection-maturation and local history can operate as rival hypotheses. These rival hypotheses must be eliminated thru axillary data sources if possible. Replicating the results of a study by a second similar study may clarify the results either by reproducing the results or by pointing out methodological errors in the first study. For example, if in the second study, the treatment group is assigned the lower group and a cross-over pattern results, this corroborates and strengthens the conclusion from the first study.

In the Increasing Treatment Effect II pattern, the treatment is assigned to the group with the lower pre-test, baseline or criterion score which increases relative to a stable control group but fails to cross-over. In this case, statistical regression operates as a rival hypothesis. If the control and the treatment groups can be tracked after the treatment and both the treatment and control groups retain the post-treatment relationship, this supports a treatment effect. If matching was possible in the initial study, replicating the study with more attention to detail in the matching step may also clarify the results.

A third pattern, the Increasing Treatment and Control Groups, occurs when both the control and treatment groups increase in the same direction but do not cross-over. In this pattern, the treatment group is assigned to the group with the highest pre-test, baseline or criterion score. The question in this result is whether the rate of increase in the treatment and control groups are the same. Unless the sample size is large or the slope of the line connecting the pre and post measurements for the treatment group is profoundly steep, a statistically significant difference in the slopes of the line connecting the scores for the treatment and control groups is unlikely. Here again, this pattern may be desirable. If this pattern is not desirable, replicating the study with an eye to matching and methodological details could provide evidence supporting a positive treatment effect.

Example 6.2

In this example, a box plot graph is used to analyze a Quasi-experimental design. In this study, new Instructor b is being evaluated against existing Instructor a. For this reason, the Instructor b is assigned the group with the highest pretest scores. If the instructors are equally effective, we would expect both instructors to maintain the pretest positions, exhibit an Increasing Treatment and Control group or an Increasing Treatment Effect I pattern.

The box portion of the Box Plots contained in Figure 6.3 were constructed by using the IQR to estimate the standard deviation for the median. The standard deviation for the median based upon the IQR is calculated using Equation 6.6. This equation assumes that the population distribution, especially is the area of the IQR, is somewhat mound shaped.

Equation 6.6: Standard Deviation Calculated from the IQR

$$S_{\tilde{x}} = \frac{1.25 \times IQR}{1.35\sqrt{N}}$$

Typically, the two sided 95% confidence interval for the mean uses the constant z = 1.96 to form a confidence interval in which p = 0.025 of the area under the normal distribution lies in each tail. However, it can be shown that when Equation 6.6 is used to estimate the standard deviation for the median, the constant z = 1.96 results in a confidence interval well beyond 99%. A compromise value of 1.7 was selected empirically and results in an approximately 95% confidence interval about the median as shown in Equation 6.7.

Equation 6.7: 95% CI for the Median

$$M \pm 1.7 \times S_{\tilde{x}} = M \pm 1.7 \times \frac{1.25}{1.35} \times \frac{IQR}{\sqrt{N}} = M \pm 1.574 \times \frac{IQR}{\sqrt{N}}$$

Figure 6.3 compares the scores for two offerings of a series of SOIs in which each series contains four SOIs. An a indicates that the SOI is delivered by the existing instructor and a b indicates that the SOI was delivered by the new instructor. For example, SOI 1b was presented to 20 students by the new instructor producing a distribution having an IQR equal to 0.89 - 0.79 = 0.10 as indicated by the black dots located on the whisker line. In the box plot, the horizontal line at the center of the box indicates the location of the median. The 95% confidence interval is created by adding and subtracting 1.574 x 0.10 / $20^{0.5}$ = 0.0352 from the median. So, the 95% upper box location equals 0.84 + 0.0352 = 0.875 and the 95% lower confidence box location equals 0.84 - 0.352 = 0.8048. In Figure 6.3, if the boxes for any pair of SOIs do not overlap, the SOIs are significantly different. In Figure 6.3, by SOI 4, the results for the new instructor are significantly lower than that for the existing instructor. This study indicates that the hypothesis that the two instructors are equivalent can be rejected.

Figure 6.3: Comparison Between Two Instructors

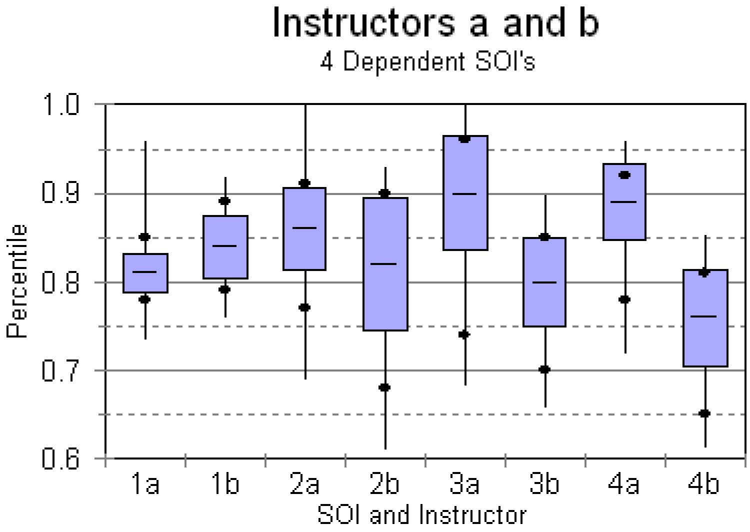


Table 6.6: Measurements Used for Figure 6.3

	1a	1b	2a	2b	3a	3b	4a	4b
Students	26	20	23	21	29	23	26	22
Max	0.96	0.92	1.00	0.93	1.00	0.90	0.96	0.86
Min	0.74	0.76	0.69	0.61	0.69	0.66	0.72	0.62
Up CI	0.83	0.88	0.91	0.90	0.96	0.85	0.93	0.81
Dn CI	0.79	0.80	0.81	0.74	0.84	0.75	0.85	0.71
75th	0.85	0.89	0.91	0.90	0.96	0.85	0.92	0.81
25th	0.78	0.79	0.77	0.68	0.74	0.70	0.78	0.65
Median	0.81	0.84	0.86	0.82	0.90	0.80	0.89	0.76

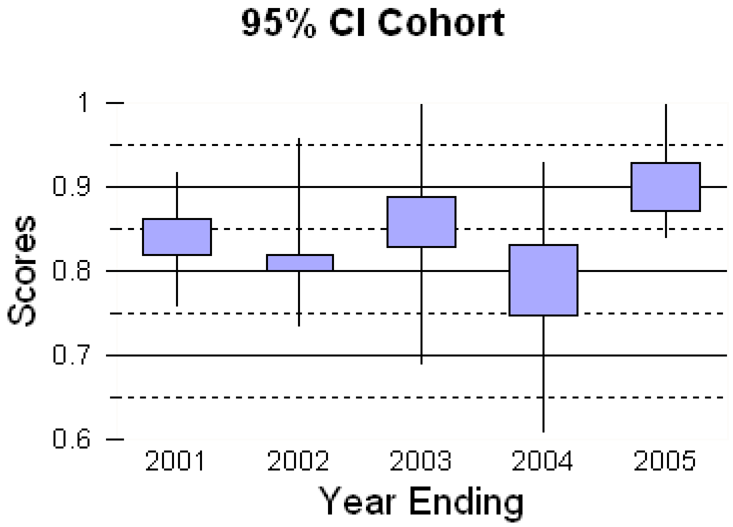
6.13 Cohort Design

The Cohort design is used in instances where the entire group of students is replaced from one instance of an SOI or collection of SOIs to the next. For example, if a college course is offered once each spring semester, the students receiving the course in a particular year form a Cohort. One of the common uses of the Cohort design is for comparison year to year under circumstances where the student population is completely replaced at regular intervals such as each year. The main requirement for the Cohort design is that the characteristics of each successive Cohort, with regard to their ability to performance successfully, changes in ways that are insignificant to the treatment condition. This is known as the assumption of quasi-comparability. This assumption is important for all quasi-experimental designs because the value of the Cohort design rests upon this assumption being met. Many elaborate methods for establishing quasi-comparability exist and a few examples are a lower limit for a standardized test required for admittance to a college or requiring prerequisite knowledge for registration in a certain course.

The Cohort design is subject to selection bias and history effects. Selection bias is the most serious and because of this, even though some method of establishing quasi-comparability is used, additional information is often collected and/or additional types of measurements are collected on each student. These are then reviewed to support the assumption that Cohorts are quasi-comparable. Due to the weakness of having only quasi-comparability to support the efficacy of the treatment, if enough students are available to form both a treatment group and a non-equivalent control group, the Non-equivalent Control Group design should be considered.

In Figure 6.4, all students completing a particular SOI during each calendar year are the treated as a Cohort. The objective is to determine if the year 2005 version of a software based SOI is an improvement over the existing instructional process which is based around an instructor delivering content. In this figure, just as with Figure 6.3, the box height represents the two sided 95% confidence interval for the median calculated using Equation 6.7. Due to the fact that the box for 2005 does not overlap those from 2001, 2002 and 2004 the hypothesis that the medians are all equivalent can be rejected. This lends support to a conclusion that year 2005 result is an improvement over three of the four preceding years. However, the 2003 offering of the SOI is not significantly different from the 2005 offering and this must be explained if possible.

Figure 6.4: Cohort Design



6.14 Concluding Remarks

The monitoring function is the most apparent indicator that an organization has a quality control program. A main purpose of this function is to show the progress made toward stated instructional goals over time. The calculations for constructing the various types of charts are not complex. However, updating charts becomes extremely tedious and expensive if done manually. For this reason, passive data collection and chart generation is very desirable.

When the monitoring function is used to the fullest extent, each measurement plotted is either a confirmation of the stability of a previous improvement effort or evidence in support of an ongoing improvement effort. Even though the results from planned studies, when successfully archived electronically, form the basis of the long-term institutional memory system, the monitoring charts are an important part of the institutional memory in the short-term. The importance of maintaining a useful institutional memory system cannot be overstated.

Often the information on certain runs charts are summarized into less detailed charts that show the organizations results over longer periods of time. These summary charts are most useful for persons at different levels in the organization. Upper management may be most concerned with year over year performance but to monitor progress, use charts containing monthly or quarterly data. At lower levels in the organization, daily, weekly and monthly charts are contain relevant time frame.

The monitoring function will also be comprised of ad hoc charts the are used to monitor progress toward specific objectives. An example of these could be the number of students exceeding the MC or TK level of performance in an SOI highlighted by reliability modeling. These types of charts are typically discontinued after the goal has been achieved and stabilized.