國立台灣師範大學教育心理學與輔導系

博士論文

指導教授: 陳柏熹 博士

Angoff 標準設定之判斷者的評估

(AN EVALUATION OF JUDGES IN AN ANGOFF STANDARD SETTING)

研究生: 張 夏石 (MICHAEL SCOTT SOMMERS)

中華民國一百六年七月

ACKOWLEDGMENTS

There are many things I've wanted to say over the years about my life here that I have never been able to say without sounding awkward or clumsy. And now, here in this place where almost no one will read them, I will finally try and put these words to rest.

This dissertation was possible only because there is a place like Taiwan. Taiwan is a place where no one thinks it's strange or unusual that you would want to read academic books for fun or conduct academic research because it just seems like the best way to spend your free time. There have been many countries created by guns and power. Many nations have been constructed through their workers, managers and the economies they have built. But the idea that there is a modern, free, and affluent society invented by men and women who value the search for, recording, and teaching of knowledge above all else is a strange idea that I think can be found today in only one place across this globe, and that place is Taiwan.

In the 20 years or so that I have lived here, I have never been refused admission to a library or forbidden from reading any books. Even some of the most expensive books in the country have been made available to me without question. Curators of private libraries and their rare contents, as well as libraries at the most prestigious public institutes of learning have never questioned my interest in their books. I have received extensive help finding books and rare collections of reading material that are now only available on microfilm. It was as if the mere fact that I wanted to know what was in their books made it impossible for the keepers of these volumes to refuse me access.

So it was that without the help of the government, the people of Taiwan, and their combined attitude toward education, research and learning, I could never have been here or done the work that you are about to read. It is to all the people that made a country where these principles are valued, perhaps like no other place in the world, that I devout this research, and where my long road of acknowledgments and my personal thanks must begin.

I have to thank the National Taiwan Normal University (NTNU), where I conducted this research, and my employer, Ming Chuan University (MCU). At NTNU, the College of Education and the Department of Educational Psychology and Counseling are part of a proud institution, and they are deservedly so. For this, I'd like to thank my advisor Dr. Chen Po-Hsi. I would also like to thank Dr. Lin Sieh-Hwa particularly for his help with and teaching of many difficult concepts in measurement and modern mental testing and Dr. Song Yao-Ting for the difference his class and tests made in my understanding of experimental design. Finally, it was only through the help and guidance of Grace Lin that I was able to manage the bureaucracy of the National Taiwan Normal University.

During the time it took to finish my classes and produce this dissertation, I taught at Ming Chuan University in Taipei and Taoyuan. My experience with the school has only been positive. For that I have to thank my directors Ada Hong and Dr. Chris Liu. It is important that I also thank Dolly Chang, as both she and Ada Hong were responsible for arranging a teaching schedule and other aspects of my job that made it easier to finish my research work at NTNU while still meeting my work responsibilities at MCU.

Many of my colleagues at MCU were helpful as friends and acted as a positive inspiration on my work. However, one of my colleagues played a very special role in my work and as such

deserves special mention. Dr. Joe Lavallee is not only my colleague and friend; he is an alumnus of the Department of Educational Psychology and Counseling. It was Dr. Lavallee who found the program and its professors. He is the first alumni of the department who entered through the pathway that I also entered. His help and inspiration throughout my studies were not only invaluable, they were irreplaceable.

A senior manager of the school that made my work possible is Dr. Nelly Chuang. At the time, Dr. Chuang was the Dean of Research and Development at Ming Chuan University that allotted the money creating the data used in this dissertation. Her assistance obtaining this money and her flexibility in allowing us to use it in the way we felt we needed it used, is greatly appreciated.

I also owe a debt of gratitude to Dr. Bao De Ming, the founder of MCU, and her son Dr. Lee Chuan, who is now the president of MCU. Their leadership and emphasis on faculty development made it possible for this research to happen in a way that it could not have in other schools.

There are many others whose help was important, not because of the academic knowledge it brought or the freedom it gave me to do the work, but because these people provided me with friendship, love, and inspiration during the years it took to produce all the different parts of this research. My daughter the little Olyvia “Cookie” Sommers is the most important person in the world to me. Every day while I worked on this project, I thought of her, I looked at her pictures, and thought of the strange and funny words that come from the mind of a 4-year-old. She is the star of my life and the little person who gives me meaning every single day: Dr. Ann Heylen has been my friend and colleague for decades, May Chen, Paul Jackson, Hans Thom, Glenn Pluckhan, Rodney Szasz, Quentin Brand, my training partners and coaches at Taiwan BJJ,

Vaughn Anderson, Professor Makoto Ogasawara, Dr. Warren Wang, Professor Andy Wang, and my family: my father Doug Sommers and my mother Sonya May, my brother Dr. Jeff Sommers and his partner Joanne Hochu, and my sister Megan. And finally, the man who saved me physically and kept me out of jail and from being deported, Hung Ming Hsieh.

Of all the named and unnamed people who contributed to my education and work, some were more important than others in helping me, leading me through serious academic troubles, while others were there for me to make sure that I stayed sane and on track. But two of my teachers were different:. Norm Cameron and Dale Beyerstein

And finally there is one point that made this dissertation special and challenging and horrible beyond belief. During the production of a dissertation things happen that are unexpected. They are not people or things that you can touch or even ideas. They can be good, and they can be bad. I suppose you could call these things luck, but since there is no such thing, a more appropriate term might just be random events. They fill your dissertation life, even as you don't know they're happening, and there's nothing you can do about them except hope like some superstitious mountain man that 'things go your way', as he rattles his bones or picks his lucky numbers. So it is with the greatest of caution I tell the writers of the billions of dissertations that are certain to follow mine, there are many things you don't want to have happen while your dissertation is whirling around you. You can get married and finish your dissertation with both love and knowledge. You can use drugs. I don't recommend this, but many dissertations have been finished under the influence of mind-altering substances. You can move your house, have babies, loose important documents or even take on too much work to really do a good job.

But never, never, never get divorced while you are writing your doctoral dissertation.

Angoff 標準設定之判斷者的評估

張 夏石 (MICHAEL SCOTT SOMMERS)

摘要

在標準設定中，專業的判斷者根據表現水準描述（Performance Level Descriptors, PLDs），扣合到標準化測驗的分數，並據以區分將學生的能力表現。這個流程通常決定了分數對學生的意義和決策人員對測驗的使用，例如，通過/未通過的決定、或優秀/平均/未通過等，也就是說，這些決定與標準設定判斷者之評估密切相關。在典型標準設定中，專家學者小組的判斷者接受訓練，評估符合表現水準的考生是否能答對測驗題目，接著互相討論判斷的結果。標準設定的組織者，則會提供回饋讓判斷者了解其決定對影響考生之通過和未通過比例的影響和其他的測驗使用情形。此外，整個標準設定過程，判斷者在訓練中被要求提出對於了解相關概念和想法之熟悉性與自信程度的自我報告，以及是否正確地來運用判斷。Angoff 標準設定是廣泛被使用於區分設定的方法之一。這個方法中，專家判斷小組對於學生的能力做出判斷，以評估學生能夠於表定時間中正確回答測驗題目。此流程相當重要，然而，有關如何地預備判斷者在標準化設定中的角色，所知仍有限。

本研究數據蒐集是由一所臺灣的大學發展之本土外語測驗和共同歐洲參考架構（Common European Framework of Reference, CEFR）所對應的題項而來，包括聽和讀兩個小組都加

以實施。本研究採用兩種共同使用的評量方法，以瞭解預備判斷者對於 Angoff 標準設定和判斷精確性的關聯。判斷的精確性是以答對率判斷的相關性(p 相關)和方均根差(Root Mean Square Error, RMSE) 和截止分數判斷（Cut-off Score Judgments, CSJ）來測量。在第一次評估時，判斷者以 PLDs 加以訓練，然後測試其對於 PLDs 切合測驗知識的 PLDs 和判斷精準性；第二次評估時，則在訓練中介紹判斷的測量精確性，對於概念和想法的熟悉性和自信程度的相關情形，發現最終判斷的測驗精確性於熟悉程度和自信程度之間沒有相關。除了主要發現之外，進一步觀察到精確的語詞說明，對於判斷的精確性是非常重要的。也觀察到以 RMSE 和 CSJ 來對精確性做出差異決定優於 p 相關。本文對未來研究方向提出在訓練 Angoff 標準設定判斷者的結論和建議，也指出本研究限制所在。

關鍵字：Angoff、判斷者、標準設定

An Evaluation of Judges in an Angoff Standard Setting

MICHAEL SCOTT SOMMERS (張 夏石）

Abstract

In a standard setting, groups of expert judges evaluate verbal descriptions of performance (Performance Level Descriptors or PLDs) contained in a standard and match these with scores on a standardized test that place students in categories of performance. This procedure is often used to make decisions about what scores mean for the students and policy makers who use the tests. For example, Pass/Fail decisions, as well as Excellent/ Average/ Fail decisions are often tied to how tests are evaluated by standard setting judges. In a typical standard setting, panels of expert judges are trained, evaluate test items, and are then given time to discuss their results with other judges. Feedback is provided by standard setting organizers that allow judges to know how their decisions would affect students Pass/ Fail rate and other decisions the test will be used to make. In addition, throughout the standard setting, judges are asked to give self-reports about their familiarity with and confidence in their understanding of the concepts and ideas during the training and whether or not the judge is applying them correctly. The Angoff standard setting method is one of the mostly widely used methods for setting cutscores. In this method, panels of expert judges make judgments about the ability of students to correctly answer test items listed one at a time. Despite the importance of this procedure, little is known about how best to prepare judges for their role as a judge in the standard setting. Data was gathered from a standard setting held at a Taiwan university to match items from a locally developed foreign language test with the Common European Framework of Reference (CEFR). The study then used an evaluation of

two commonly used methods to prepare judges for an Angoff standard setting and their relationship with judge accuracy. Both a listening and reading panel were conducted. Accuracy of judges was measured by the p-value correlation, the Root Mean Square Error (RMSE), and the Cutoff Score Judgment (CSJ). For the first evaluation, judges were trained in the PLDs and then tested about their ability to match a test of knowledge of the PLDs with the three measures of judge accuracy. No relationship was found between tested knowledge of the PLDs and judge accuracy. The second evaluation correlated familiarity with and confidence in the concepts and ideas introduced during the training period with the measured accuracy of the judge. Once again no relationship was found between familiarity and confidence with the final measured accuracy of the judge. In addition to the main findings, it was also observed that the exact wording of the instructions to instructions is very important to the accuracy of the judges. RMSE and CSJ were observed to make different decisions about accuracy than the p-value correlation. Future directions for research on the training of Angoff standard setting judges are suggested, as are the limitations of this study.

Keywords: Angoff, judges, standard setting

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1 INTRODUCTION

1.1 Significance of the Current Study

This research is about the training and preparation of judges for the Angoff method of standard setting. In particular, its goal is to breach a hole in the current research concerning the efficacy of training and procedures for the method. There currently exists little information that addresses questions about how the training and procedures affect the performance of Angoff judges in a standard setting.

Sometime during China's Han Dynasty, it was discovered that a sample of someone's knowledge about a particular subject could be used to estimate the total amount of knowledge known by that person on that subject. This was the discovery that became what we now call 'the test' (Elman, 2000). Over thousands of years, testing has spread across the globe and its use become ubiquitous in selecting the most suitable. Despite this, its design has remained largely unchanged for almost all of those thousands of years. It was not until the 1960s and 70s that the question of what a score means about the test-taker placed demands on the test that it could not yet handle (Glaser, 1963). With this pressure came the realization that understanding of this new idea of testing lagged far behind the actual practice. The decades that followed, the 1990s and 2000s, saw an explosion in work on this problem, and standard setting became the established method for determining how the scores on a test would be understood. This procedure came to be a key aspect of what is now called criterion-referenced testing.

The design of a criterion-referenced test has now become highly standardized. Manuals, and standard designs dominate training, procedures, and materials (Egan et al., 2012; Loomis, 2012) through national organizations and guidelines that define how a 'quality' standard setting is

conducted. The current complexity with which standard setting operates leaves an observer feeling a high level of ability has been obtained, and that standard setting, is an advanced procedure. Sophisticated methods in standard setting, such as vertically-moderated standard setting (Huynh & Schneider, 2005; Lissitz & Huynh, 2003; Lissitz & Wei, 2008) are now used to produce results (Raymong & Reid, 2001).

Despite this, many aspects of the standard setting have yet to be explored. Very little is known about the implications of the training and procedures of standard setting judges. All standard setting methods call for training and procedures and some of these are quite complex. It seems strange to say that for the most widely-used, there is little understanding about the ways in which training methods and procedures prepare participants in the standard setting to perform their tasks, and that the reasons to believe that someone has been adequately trained to set the standard are drawn largely from their face validity (Holden, 2010).

The purpose of this study is thus to examine the relationship between conventional training procedures for the Angoff standard setting method and the final outcome of the standard setting. The idea that training and the procedures of the method should have a positive effect on judges' ability is almost too obvious to state. Yet because of the lack of real data on the subject matter, it is not entirely clear that this is true and in what way it could be true. The study that follows is an attempt to clarify this with empirical data drawn from an actual operational Angoff method of standard setting.

1.1 Research Questions

The analysis conducted in this study seeks to address the following research questions:

- Does knowledge and training in Performance Level Descriptors (PLDs) work effectively to predict an Angoff standard setting judge's ability?
- Do self-report measures of familiarity and confidence with one's knowledge of procedures and materials work effectively to predict an Angoff standard setting judge's ability?

1.3 Terminology

PLD – The abbreviation for Performance Level Descriptor. PLDs are verbal descriptions of what a candidate can, and sometimes cannot, do at a particular score on a given test.

Social Influence – Judges in a standard setting may be affected by a range of factors. Social influences refer to those influences that originate in the personality and individual differences between judges. These may vary from judge to judge, in contrast to influences that are, for example, demographic or procedural.

Familiarity – Familiarity refers to the degree to which a judge feels his or her exposure to the concept have made it understandable. This is often measured with a Likert-type scale that asks judges to indicate how much they believe their evaluation is an accurate evaluation of the rating of a particular characteristic. In this study, familiarity is measured by a series of self-report Likert-type surveys.

Confidence – This is a concept related to familiarity. Confidence refers to the degree to which a judge believes that his or her decisions are correct. This is often measured with a Likert-type scale that asks judges to indicate how much they believe their evaluation is an accurate evaluation of the rating of a particular characteristic. In this study, confidence is measured by a series of self-report Likert-type surveys.

Accuracy – Accuracy refers to how close a judge's judgment is to the actual value of something. In this study, judges are asked to estimate a cutscore for items used in an Angoff standard setting. The accuracy of their judgments are assessed using two different measures, the p-value correlation and the Root Mean Square of the Error (RMSE).

CHAPTER 2 LITERATURE REVIEW

2.1 Standard Setting Method

This section will review some important aspects of the standard setting procedure, and some of the identified problems that make it difficult to interpret the meaning of standard setting scores.

Standard setting refers to the family of procedures used to establish cutscores on a scaled examination. Cutscores separate scaled scores into categories of performance defined in a performance standard (Cizek, 1996; Cizek, 2001; Cizek & Bunch, 2007; Cizek, Bunch & Koons, 2004). Standard setting is mostly used in criterion-referenced examinations to match standardized test scores with a verbal description defined in performance level descriptors (PLDs) of the performance standards. Panels of judges use different methods to compare PLDs with different types of information about items or examinees. The term "standard setting" is used to refer to the different procedures and materials used to make these cutscore decisions. Since the first suggestion of this idea in the 1950s (Nedelsky, 1954), dozens of different procedures have been developed. In one survey (Kaftandjieva, 2010), more than 60 different methods were identified with more than 15 appearing since the year 2000.

Standard setting grew out of the expanded role of “criteria” in testing. Examinations can be defined as norm-referenced or criterion-referenced (Glaser, 1963; Shepard, 1980). Norm-referenced tests produce results that allow for comparison between individuals and dominated high stakes testing for much of the last century. Such tests are limited by an inability to indicate what the score means for examinee ability. Criterion-referenced tests produce results that have assigned a defined meaning to a particular score. These abilities are typically defined in

descriptions ranking them from least to most capable. Such descriptions are referred to as a 'performance standard' and the descriptions that define individual categories of performance as 'performance level descriptors' or PLDs. The standard setting allows for these ranked descriptions - the PLDs - to be placed along scaled test scores providing latent trait scores that correspond with the different categories of ability defined in the standard. Cizek and Bunch (2007, p. 13) have stated that,

Standard setting refers to the process of establishing one or more cutscores on a test...Cutscores function to separate a test score scale into two or more regions, creating categories of performance or classifications of examinees.

A large number of different standard setting procedures have been developed (Hambleton & Patoniak, 2006; Kaftandjieva, 2010; Cizek & Bunch, 2007). While these procedures vary enormously in their details, they all share one property. These procedures present panels of trained experts (the judges) with performance standards and different types of information about items and examinees. The judges are then asked to use these procedures to decide what score on the test is the cutoff point between the different categories of performance. The actual procedures used can vary considerably and different procedures may use a wide range of different types of information. A typical convention in contemporary standard setting procedures is to permit a significant amount of input to inform judges about the impact of their cutscore decisions. For example, one common way to handle this is for panel organizers to allow discussion between judges about their decisions, and then tell them what percentage of an actual examinee population would fall above and below their cutscore decisions.

As a result of this wide range of methods and procedures, different panels do not always agree on the cutscore decision, even for the same test items and with the provision of the same feedback information about pass/fail rates. It has long been known that different methods produce estimates of cutscore decisions that are systematically different based on their differing procedures (Buckendahl et al., 2002; Green et al., 2003; Hambleton & Patoniak, 2006; Reckase, 2006; Yin & Schultz, 2005). Even small changes in standard setting procedures can result in changes in judge's decisions (Cross et al., 1984; Hertz & Chinn; 2002; Jaeger 1982). Judges, or even the same judge, may not make the same judgments under apparently identical conditions (George, Haque & Oyebode, 2006).

Very little has been written about the validity of the various standard setting procedures. The concept of 'validity' is itself a complex and contested issue. Many different definitions have been suggested. The National Council on Measurement in Education (NCME, 2015) defines it as, "…a general term used to describe whether or not the interpretation of a theory is plausible." One widely cited definition (Kane, 2006) attempts to explain the complexities of the term.

Measurement uses limited samples of observations to draw general and abstract conclusions about persons and other units (e.g., classes, schools). To validate an interpretation or use of measurement is to evaluate the rationale, or argument, for the purposes being made…Ultimately, the need for validation derives from the scientific and social requirement that public claims and decisions be justified. (2006, p.17)

This is similar to the definition used in the S*tandards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014, p. 9) that suggests, "Validation can be viewed as developing a

scientifically sound validity argument to support the intended interpretation of test scores and their relevance to the proposed use."

In contemporary psychological testing, a general theory of validity, sometimes referred to as the argument-based concept of construct validity, has emerged as the dominant model (Cronbach, 1988; Cronbach & Meehl, 1955; Kane, 2006; Loevinger, 1957; Messick, 1981, 1989, 1998). An argument-based concept of validity,

…first lays out a network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the performance (Kane, 2006, p. 23).

Because the structure of a standard setting is so different from that of an experiment or correlational study, the discourse about validity in standard setting is quite different and separate from these more mainstream models of psychological and educational testing. Building on the definitions for standard setting validity suggested by Cizek, Kane and other modern standard setting theorists reject the conceptualization of standard setting as a psychometric technique with knowable or estimable parameters (Cizek & Bunch, 2007, p. 18). Given Cizek's (2001a) belief that cutscores are arbitrary and that their importance comes from their usefulness rather than their validity, Kane (2001) stresses the point that procedural evidence is the most significant aspect when validating a standard setting, in other words, was the standard setting done the way it was supposed to be done. Examinations of validity in standard setting methods appear to be based on a series of *ad hoc* principles (Kane, 1992, 2001) and derived from the approach that accepts "just because a standard setting is arbitrary does not mean it is not useful" (Hambleton, 1980, p. 102).

Following in this tradition, Hambleton (2001; see also Schafer, 2005) built on this suggestion that further information is necessary to determine the 'usefulness' of the standard setting, Kane (2001, p. 63) states that,

Procedural evidence is especially important in evaluating the appropriateness of performance standards. In most cases, few if any solid empirical checks on the performance standards are available.

So rather than the conventional forms of validity defined for conventional psychology, a different model of validity has been suggested for standard setting. This includes (1) a Definitional perspective - that "To be called performance standards, there must be operationally defined, mutually, exclusive, exhaustive ordered categories and a decision process based on one or more assessments to place tested subjects in those categories", (2) a Psychometric perspective - that "…form a scale that can be evaluated using the well-known criteria of reliability, validity, and utility" (3) a Legal perspective – that "Performance standards are a part of a decision-making process. Assuming the decisions have importance (i.e., stakes, which imply the possibility of harm), the process may be held to criteria that courts have determined are crucial for legal acceptability", and (4) an Institutional perspective – that a standard setting must be consistent with the goals of the institution sponsoring it. These points should also cover the legal defensibility, the generation of assets and the efficient uses of resources in the construction and use of the standard (Kane, 2001; Schafer, 2005, p. 62).

Cizek and Bunch (2007) suggest that panel organizers should report a number of statistical tests to support their argument for validity. In contrast, Dixy McGinty (2005) has pointed out that such statistical tests as these are really more accurately thought of as indicators of reliability, and

while useful in demonstrating validity, are not themselves measures of validity. As a result of this confusion, in comparison with other psychological assessment procedures, a scientific justification for the validity of a particular procedural decision, such as choice of a method, or variation in a procedure, is very rarely given and when this is done, such justifications are typically operational. Given this definition of validity for a standard setting, explaining choice of a method or variation in a procedure may not be possible or even necessary.

It is widely stated that standard setting procedures are dominated by two methods that are historically linked - the modified-Angoff method and the Bookmark method (Cizek & Bunch, 2007; Engelhard, 2007). The modified-Angoff method is derived from an original method named after William Angoff who, ironically only briefly mentioned it as a note, and attributed the idea named after him to his colleague Ledyard Tucker (Cizek & Bunch, 2007). The main principle of the method is that items are examined one-at-a-time and judged in various ways for their suitability to make decisions about examinee classification. Since the Angoff method is the main focus of this study, much more will be said about it in the following sections; however, the Angoff method is widely cited as being, "the most commonly used method for setting performance standards in contemporary use in licensure and certification context" (Cizek & Bunch, 2007, p. 82). Regardless of the literal accuracy of this statement, it is unquestionably a widely used method to produce cutscore decisions for high stakes tests.

The other widely used standard setting method is the Bookmark method. The Bookmark method emerged from procedural difficulties with the Angoff method. It was first suggested by Mitzel et al. (2001), although Cizek and Bunch (2007) trace its roots back to procedures extended from the Angoff method and used in the 1990s by researchers at American College Testing (ACT) for the

National Assessment of Educational Progress (NAEP). In the Bookmark method, items are placed in a booklet, referred to as the Ordered Item Booklet (OIB), where they are ranked by their difficulty measures. Judges then place a marker on the item that separates the various categories of the performance standards. Engelhard (2007) speculates that, because of its widespread use in assessments related to the American educational policy No Child Left Behind (NCLB), the Bookmark method may have been the most widely used standard setting method.

Standard setting is now a routine aspect of test development. Huge numbers of the procedures are performed regularly during the development of state and private tests. Standard setting panels were conducted as part of the No Child Left Behind (NCLB) network of accountability tests used in the USA (Linn, Baker, & Betebenner, 2002; Linn, 2003), as well as in other public education accountability projects throughout the world. Standard setting also plays a role in the development of the examinations that establish standards for a wide variety of occupations and professions (Nelson, 1994). In addition, panels similar to those in the standard setting are increasingly used for other purposes. For example, Roach, McGrath, Wixson & Talapatra (2010) describe a procedure similar to a standard setting panel to 'align' two or more different types of assessments whose content is not directly comparable. The results of their study resemble what could be produced from a mathematical equating of different assessment procedures. Their application of the panel comparison, instead of an equating, stems from the limited use of the assessments and hence limited numbers of observations available to perform an equating.

2.2 The Angoff Method

This study deals specifically with the Angoff method of standard setting. The section that follows briefly describes its history and the procedural aspects of the method that distinguish it from other methods of standard setting.

As mentioned above, the Angoff method is named after William Angoff (Angoff, 1971) who, attributed the source of the method named after him to his colleague Ledyard Tucker (Cizek & Bunch, 2007). The Angoff standard setting method is one of the oldest methods and is reputed to be among the most widely used methods in the world for setting cutscores (Cizek & Bunch, 2007). From a research point of view, the Angoff method is particularly useful because it produces many discrete values at points throughout the procedure, permitting the application of techniques derived from classical, as well as latent trait theories, such as Item Response Theory (Embretson & Reise, 2000) and Rasch modeling (Bond & Fox, 2001).

There are many different versions of the Angoff method in use today. For this reason, methods that belong to the Angoff family of standard setting methods are sometimes described as a "modified-Angoff". It has been suggested that there is no general agreement on a definition of the Angoff method (Brandon, 2004; Reckase, 2000), although Brandon (2004) lists 5 steps he believes characterize the modified-Angoff procedure,

1. selecting judges
2. training judges
3. defining and describing the performance level descriptors
4. estimating examinee performance at the level of each item

5. review of empirical information by judges and discussion of item estimates

This definition, while widely cited, is difficult to use. All of these points are routine aspects of other standard setting methods and only number (4) is an aspect distinctive to the Angoff family of standard setting methods. While estimation of examinee performance at the item level is found in other methods, such as the Nedelvsky method (Nedelvsky, 1954) the way it is done in the Angoff methods offers a true distinction between the modified-Angoff and other standard setting methods.

The modified-Angoff is distinctive in its procedures for estimating the cutscore in that,

1. Judges are presented with items one-at-a-time.
2. Judges are asked to estimate examinee's ability to answer the item correctly.
3. Estimation of examinee ability to correctly answer the item is done item-by-item, and items are not necessarily presented in any particular order.

The second point, estimation of an examinee's ability to answer the item correctly, has been done in many different ways. Brandon (2004, p. 60 note 2) provides a partial list of some of these different ways. Sometimes percentages are recorded instead of probabilities. Sometimes judges specify the number of candidates out of 100 who could answer the problem correctly. (e.g., Engelhard & Anderson, 1998; Impara & Plake, 1998). Sometimes judges are given a choice of range of percentages or proportions. For example, Cross, Impara, Frary and Jaeger (1984) and Plake and Giraud (1998) instructed judges to select from deciles. Halpin, Sigmon and Halpin (1983) printed the lowest acceptable probability and the highest probability. Cizek and Bunch (2007) list several different versions of the modified-Angoff, including the yes/ no Angoff

procedure in which judges indicate only a yes or a no concerning their judgment of examinee ability to answer the item correctly.

In addition, the modified-Angoff standard settings conventionally incorporate a number of other procedures to produce a convergence of scores across judges. These are referred to in number (5) of Brandon's (2004) list, and include,

1. Judges have several opportunities to refine their estimations, referred to as 'rounds'. The current convention is to perform a standard setting in sometimes two, but often three, rounds (Cizek & Bunch, 2007).

2. In between rounds, judges have the opportunity to compare their estimations with each other and discuss why they made their individual decisions. This is referred to as 'discussion' (Cizek & Bunch, 2007).

3. In addition to discussion, judges are presented with data that reflects the impact of their decisions. For example, judges may be shown the percentage of examinees who would fall above or below their estimated cutscores. This is referred to as 'impact data' or 'feedback'.

Virtually all standard settings, no matter which method is used, when conducted with multiple rounds, discussion among judges between rounds and the provision of feedback demonstrate a convergence of judges' cutscore decisions across rounds. So characteristic is this result that Cizek (2001a, p. 10) refers to it as a "common feature of standard settings". This convergence is not unanticipated. Experts, given the opportunity to discuss data relevant to their expertise, will develop elaborate explanations for the data based on information drawn from their shared background (Chi, Glaser & Farr, 1988; Johnson, 1988; Larkin, McDermott, Simon & Simon,

1980). It is thus reasonable to interpret the convergence of cutscore decisions as a growing expert consensus about the contents of the standard setting and its panels.

However, the exact origin of these effects is not well understood and much discussion has been generated about their origin. Many types of effects have been suggested as a potential issue in the converging scores of the judges. For example, some standard setting literature has examined social influences during the discussion drive cutscores toward agreement (Fitzpatrick, 1989; Hurtz & Auerbach, 2003; Hertz & Chinn, 2002; Wessen, 2010). Social influences, such as the effects of dominant individuals or group conformity, may be driving judges to report cutscore decisions that are more and more similar to each other. The mechanism of these biases has not been well-established. Despite widespread speculation about the role of these social influences (Fitzpatrick, 1989) and some empirical examinations (Hurtz & Auerbach, 2003; Hertz & Chinn, 2002; Wessen, 2010).

Little is really understood about the social influences on the standard setting. Attempts to measure them have been largely unsuccessful. All of them seem to be derived from an ad hoc ‘common sense’ idea of what effects could be operating in the standard setting rather than from a theoretical description of what kind of factors exist in the procedure. In fact, it is not even clearly understood how they could operate or even if they exist in a fashion that would affect the outcome of the standard setting. As a result, there continues to be confusion about how judge’s accuracy is influenced or even what could be influencing it.

2.3 Training and the Angoff Standard Setting Method

Merriam-Webster tells us that training is, "a process by which someone is taught the skills that are needed for an art, profession, or a job." This definition implies both learners and teachers. All standard setting methods would have both. But in addition, the idea of 'training' implies that it is preparing learners for a task that they cannot or do not perform naturally without instruction. In principle, learners who perform well during training should be better prepared for the task than those who do not perform as well.

The training that has emerged for the Angoff method of standard setting has two fundamentally different manifestations. In the United States, where most of the published research on standard setting is generated, training has focused on preparing judges to identify as accurately as possible where categories of the borderline student test takers end and begin - or at least this has become the emphasis of activities used during standard setting training. In Europe, on the other hand, the Common European Framework of Reference (CEFR) has for years held the dominant position as a standard in language testing. The CEFR is fundamentally nothing more than lists of the PLDs that are aimed at describing competency. As such, training for standard setting that has emerged from Europe is based largely on whether judges are able to use and perform tasks based in lists of the skills associated with different levels of competency.

The American sense of standard setting is very clearly laid out around the concept that the judge's job is to identify the borderline test taker. Many researchers describe in detail the tasks that such a student should be able to do. For example this quote from the widely cited Raymond & Reid (2001, p. 147) illustrates this point.

Training should give participants an opportunity to practice the steps for assigning MPLs (Minimum Passing Level) under conditions similar to the conditions they will experience when assigning actual MPLs (p. 144). Asking the participant with the lowest MPL and the participant with the highest MPL to explain their rationale is a common training technique.

Similar descriptions can be found in more recent examples of the training of standard setters. Raymond & Reid (2001, pp. 150-1) provide explanations of a training program for standard setting judges. Loomis (2012) provides details of the preparation of standard setters for NAEP. While she gives descriptions of how it is that NAEP selects and prepares their standard setters, both her work and that of Raymond & Reid (2001) fail to provide any evidence that their training methods can actually produce the knowledge and skills deemed necessary by the authors. It is not at all clear that asking judges to explain their reasoning has any effect on their actual ability to perform the task with more or less efficacy. Although such instructions seem to make sense, to do so, in effect, is relying on the face validity of the activities.

In other disciplines and areas of psychology, this form of work would be done through the use of clinical trials to assess the efficacy of training methods. Instead, judges are asked to fill out self-report surveys describing self-reflections on their knowledge and feelings about training and personal success at mastering the training. Gregory Cizek (2001, 2012, 2012a; Cizek & Bunch, 2007) is one of the leading researchers in standard setting today. He has discussed in great detail the use of these self-report surveys to monitor progress during the standard setting and to clarify judges' level of knowledge and attitudes during and after the procedure. While Cizek has published several major academic books on standard setting, they are better thought of as manuals concerning how to conduct a valid standard setting. Cizek (2012a, p.170) states that,

Minimally, two essential validity related questions are addressed by the surveys. (a) Is there evidence that the standard setting participants received appropriate and effective training in the standard setting method, key conceptualization, and data sources to be used in the procedure? (b) Is there evidence that the participants believe they were able to complete the process successfully; yielding recommended cutscores that they believe can be implemented as valid and appropriate demarcations of the relevant categories.

Cizek (2012a) continues by providing extremely detailed examples of 'evaluations' timed for the "End of Orientation" (p. 174), the "End of Method Training Session" (p. 175), the "End of Round One" (p. 175), "Round Two" (p. 176), "Round Three" (p. 177), the "Final Evaluation" (p. 177), and a final form dealing with "Level of Reliance on Information" (p. 178).

The study reported here was originally planned long before Cizek (2012a) wrote this, but in addition, it has a different agenda in mind. As such, its schedule only roughly follows the one suggested by Cizek (2012a). Of the seven different types of assessment used in this study, five of them were self-report surveys; addressing the,

1. knowledge to standard setting procedures
2. knowledge of the CEFR
3. knowledge of the Practical English Test
4. beginning of Day 2, prior to the beginning of the operational standard setting
5. final evaluation

In one final suggestion for training, Loomis (2012), and also Cizek (2001), describe the slightly different version of this used by NAEP. NAEP uses as a key element, and "the most essential

part of the process", (Loomis, 2012, p. 123), the concept that the training process should produce in judges a common understanding of the achievement levels. As a result, one of the procedures is that all judges take a version of the test to try to understand what it will be like for the actual test takers. A similar situation exists in Europe. European standard setting of language tests is based largely around the Common European Framework of Reference (CEFR). Some CEFR manuals are simply lists of PLDs for various language situations (Council of Europe, 2001). Others are lists of PLDs and how to interpret them. Exercises for the training of judges can be constructed from these instructions (Council of Europe, 2001; Council of Europe, 2009). Some of these exercises are very interesting and appear to have very strong face validity. But like their American counterparts, a formal test of their ability to predict standard setting outcome has yet to be reported.

The situation described here, where the training for judges is suggested without any quality control other than the face validity of the procedure, is in fact much more significant than first indicated. It is difficult to find any source anywhere dealing with the training of standard setting judges that reports or even describes a need for predictive checks on training activities. Hambleton & Pitoniak (2006) spend a great deal of time in their chapter "Setting Performance Standards in the APA Publication Educational Measurement" discussing the importance of training. Many standards of the APA's *Standard for Educational and Psychological Testing* are cited in Hambleton & Pitonik (2006) as the authors refer to the training of judges. For example, on page 434, they state,

Standard 4.20 addresses the desirability of obtaining external evidence to support the validity of test score interpretations associated with performance category descriptions.

Standard 4.21 stresses the importance of designing where panelists can optimally use the knowledge that they have to influence the process.

The paper itself contains numerous sections that deal directly with the training of judges, such as "2.4 Step 4: Train Panelists to Use the Method" (p. 437) and "6. Training Panelists" (p. 453-455) which are cited extensively in Raymond & Reid (2001). At no point in any of these references is there mention of an empirical validation of these training methods or how such training methods are connected to the standard setting method in a way that makes them more valid as methods of performing the procedure.

2.4 Problems with the Angoff Method

The Angoff method suffers from all of the general problems that plague standard setting, such as the rejection of the conceptualization of standard setting as a psychometric technique capable of discovering a knowable or estimable parameter (Cizek & Bunch, 2007) rather than an abstraction that is useful, but not a real value (Cizek, 1996; Cizek, 2001; Cizek & Bunch, 2007; Cizek, Bunch & Koons, 2004). Or as Dixy McGinty (2005) put it, the way in which statistics describing standard setting performance are not really indicators of validity, and while useful in demonstrating it, are not themselves measures of that validity.

But also, in the Angoff standard setting, there are a series of special problems that judges experience. The most prominent of these are related to the fact that Angoff judges are making estimates of the difficulty of the items. Human minds are limited in their ability to make such estimates (Brandon, 2004; Goodwin, 1999; Impara & Plake, 1998; Linn & Shepard, 1997, Lorge & Krulou, 1953;; Norcini et al., 1987; Norcini, Shea & Kanya, 1988; Shepard, 1994; Smith & Smith 1988; Taube, 1997) and the fact that the Angoff method forces judges to do so as part of

the procedure produces a situation in which judges look for extra sources of information on which to make their estimates, such as the copying estimates of other judge's estimates, or the copying of feedback information, and the problem of restricted range.

Copying of Judge's Estimates and Feedback Information

One source of information that may be producing convergence of the p-values is the provision of feedback information. This phenomenon has been widely studied, but its root cause is not clearly understood. It is generally seen as an indication of improving performance of the judges as they receive more information about the items (Cizek, 1996; Cizek, 2001; Cizek & Bunch, 2007; Cizek, Bunch & Koons, 2004). A second explanation which will be discussed later is that the judges are simply copying the information they are receiving about the items during the feedback sessions. In its modified form, the Angoff standard setting method asks judges to rate the difficulty of test items. The ability of expert judges to make such estimates is crucial to the validity of the method, and as such, a large and comprehensive research literature has been developed to address the issue.

P-value correlations as high as those seen in later rounds of an Angoff method standard setting are virtually never seen unless judges have not already been told the p-value of the items during feedback information before Round 2 and 3. A large number of references are typically cited questioning the ability of even the most highly trained experts to accurately estimate the difficulty of test items and the way in which asking judges to make an estimate has an effect on the magnitude of the estimate (Brandon, 2004; Goodwin, 1999; Impara & Plake, 1998; Lorge & Krulou, 1953; Linn & Shepard, 1997; Norcini et al., 1987; Norcini, Shea & Kanya, 1988; Shepard, 1994; Smith & Smith 1988; Taube, 1997).

Angoff method procedures conducted by Brandon (2004) concluded that typically, the values obtained by correlating the empirical p-values with the estimates obtained from Angoff judges range from around 0.40 to 0.70, indicating that at best, actual estimation of the p-value can rarely account for more than half of the variance in a judged estimate. In conclusion, he states (p. 71), "results of this level, show that the ordering of item estimates - particularly those in operational standard setting studies - can be expected to mirror moderately the ordering of item difficulty."

The clustering of scores around a central point is referred to as 'restricted range'. And if these scores are clustered around the middle of the rating scale, this is referred to as 'centrality' (Saal et al., 1980). One indication of this would be that estimated values for the difficulty of items that suffer from centrality would have a smaller standard deviation than the standard deviation of the measured items, indicating that estimated values for the easy items and for difficult items are not correct (Saal et al., 1980), and are more correct for items closer to the mean or median. This, in fact, is a commonly observed aspect of the research. Lavallee (2012, p. 14) reviewed the literature related to this issue and concluded, "…results consistent with a centrality effect have been found *every time they have been looked for*" (italics in the original). In addition, the tendency for judges to cluster estimates of actual values in tighter distributions than the actual values themselves has been the subject of comment for almost as long as there has been systematic scientific investigation into standard setting results.

Lorge and Kruglov's original (1953) study found a standard deviation of 16.3 for the judges' estimates compared with 23.7 for the empirical difficulty values. Goodwin (1999) reported, in her study of the results of a financial planner licensing exam, that the judges' estimated p-values were "more homogeneous" than the actual results obtained from the administration of the items to candidates. The standard deviations for the estimates of total group and for borderline

examinees were .09 and .10 respectively. The actual observed values were.19 and .18.Van de Watering & van der Rijt (2006) compared the estimates of difficulty values for teachers and students. They found high rates of inaccuracy among these groups. Interestingly, their student group did not overestimate the difficulty of easy items, although they showed dramatic underestimation of difficult items. Teacher's estimates of easy items showed much more centrality and systematically underestimated the easiest items.

More recently, Brian Clauser and his team have expanded on this theme with attempts to find out more about what and how different factors drive p-value estimates. Clauser et al. (2013) confirmed that providing judges with information about the empirical p-values of items was what resulted in the characteristic distribution of judge's estimates in a standard setting. Mee et al. (2013) found that varying the instructions judges received could also affect their final cutscore. Clauser et al. (2014), using a generalizability theory framework, found that the greatest source of variability within the Angoff standard setting was between tables rather than between individual judges, suggesting that something similar to social influences could be effecting estimates inside the tables of the judges.

CHAPTER 3 METHODS

3.1 Materials

The data used in this study is drawn from a standard setting meetings held at a Taiwan university (hereafter referred to as The University) to link a university-level English proficiency exam to the Common European Framework of Reference (CEFR) (See Appendix 1). The test used in this study, the English Proficiency Test or EPT, is an examination of English as a Foreign Language. The EPT is a series of in-house language proficiency tests developed to meet the needs of the Practical English (PE) program adopted at The University. The EPT exams are multiple-choice exams. They test a series of listening, reading and vocabulary skills in a number of different practical contexts. It is divided into 8 sections with students' progressing through 2 sections of Practical English each year: PE 1 and 2 are taught to freshmen (1st year), PE 3 and 4 to sophomores (2nd year), PE 5 and 6 to juniors (3rd year), and PE 7 and 8 to graduating seniors (4th year). An outline of the test organization is detailed in Table 3.1 (Lavalle, 2012). Items on the EPT are linked to vocabulary suggestions contained in the PE textbooks. The CEFR was not taken into account for items used in this standard setting, which are tied to topics covered in the textbook, rather than the CEFR scales. The textbooks from The University were not designed with the CEFR in mind; however this is one of the goals of the standard setting, to match the textbooks and items written within the PE program with the standards of the CEFR.

Because the same items may appear on more than one PE test, The University maintains a strict control policy over them. As a result, no examples of items can be provided in this research.

Items for the EPT are written by the classroom teachers of the PE program under the supervision of test editors who are assigned by the school. Items are then sent to a proofreader and finally

returned to the editors. The test editor returns the test to the school who then print and distribute the test forms to students. The various tests of the EPT are administered on a single day. So for example, all freshman students receive the test at the same time. All sophomores receive the test at the same time, which is different from the time for freshman students and other students.

Following student examinations, test results are collated, sent to a test coordinator and calibrated with Winsteps Rasch modeling software (Linacre, 2012). All test items are placed on a single difficulty scale. Items are sorted by their point-biserial correlations and difficulty values, and stored in an item bank for later use. Currently, most items that appear on the EPT are drawn from this item bank, although teachers continue to write new items to expand the item bank.

The test items used in this standard setting were drawn from several different midterm examinations. All items had been calibrated onto a single scale using Rasch modeling. This standard setting project was designed to establish cutscores along the scale used to calibrate all items in the item bank and not along a raw score scale corresponding to a single test form. Accordingly, the test form used in the project was actually a composite, with its items drawn from a series of different test forms administered during the midterm examination period for first-, second-, third- and fourth-year students. The tests shared a number of common items illustrated in Table 3.1, which were used to equate them and calibrate them together onto the same scale.

Table 3.1.

Contents of the English Proficiency Test (EPT)

Skill	Question Type	Description	Items	Time
L	What's next?	Student hears 2 conversational turns and is asked to choose the next response.	20	~45 min
	Dialogues	Student hears short conversation of about 8-14 turns and answers 3-5 comprehension questions.	10	
	Extended Listening	Student hears a short monologue and answers 3-5 comprehension questions.	15	
Total			45	
R	Fill in the Blank	Student chooses a word or short phrase to complete a sentence.	10	~55 min
	Cloze Reading	Student chooses words or short phrases to complete a short passage (multiple-choice cloze).	10	
	Reading with Questions	Student reads a short passage (150-300 words) and answer 3-5 comprehension questions based on the text.	30	
Total			50	
TOTAL			95	100 min

3.2 Judges

Judges were selected primarily from faculty and staff of The University. Several external judges were selected to provide diversity to the standard setting decisions. These judges were selected because of their experience teaching students at similar universities in Taiwan. Two of the external judges were faculty members at universities in the Taipei area and one was a doctoral candidate at another university but had taught remedial classes for the university at which she was studying. Table 3.2 (Lavalle, 2012) provides a summary of the judges and a brief description of the background of each.

Table 3.2.

Angoff Judges

Panel	Judge	Gender	English NS/NNS	Position
1 (Mon)	Agf11	F	NNS	Administrator, former teacher
	Agf12	F	NNS	Teaching Assistant, recently graduated student
	Agf13	F	NNS	Teacher
	Agf14	F	NS	Teacher
	Agf15	M	NNS	Teacher
	Agf16	F	NNS	Teacher
2 (Wed)	Agf21	M	NS	Teacher
	Agf22	M	NS	Teacher, External University
	Agf23	F	NNS	Teacher
	Agf24	M	NNS	Teacher
	Agf25	M	NS	Teacher, External University
	Agf26	F	NNS	Teaching Assistant,
3 (Fri)	Agf31	F	NNS	Teacher, External University
	Agf32	F	NNS	Teacher
	Agf33	F	NNS	Administrator, Teacher
	Agf34	F	NNS	Administrator, Teacher
	Agf35	F	NNS	Teacher
	Agf36	F	NNS	Teacher

F=female, M = male, NS = native English speaker, NNS = non-native English speaker

3.3. Procedures

A one-day training/orientation session was held on Saturday, July 10, 2010 for all the participants. The judges themselves were then separated into three different panels which were held on Monday, July 12, Wednesday, July 14, and Friday, July 16 in 2010. The individual panels were conducted on three separate days to ensure that proper procedures were followed, particularly during the discussion period. A group of six judges participated on each day. The facilitator for each discussion session acted as the moderator of each of the panels, thus requiring having the panels meet on separate days.

Introduction to Training

As noted in Table 3.1, the test form presented to each of the panels was a composite drawn from tests in the EPT series of tests. The items were drawn from test forms administered as part of the annual EPTs for all four year levels of the program, and differed slightly from the EPT exam described earlier. Table 3.3 (Lavalle, 2012) summarizes the type of question types used in the composite form that each of the judges had to work with. The form itself was composed of a listening and a reading section.

Table 3.3.

Contents of the Test Form Used in the Standard Setting

Listening		
	What's Next?	16 items
	Dialogues	12 items (3 listening texts)
	Extended Listening	12 items (3 listening texts)
Reading		
	Fill in the Blank	10 items
	Text Completion	16 items
	Reading with Questions	14 items

For the purposes of acclimatizing judges to difficulties encountered taking the test and provide them with the experience of taking the exam, a training form was created with the same format as the regular exam (Loomis, 2012). The test form used in the operational standard setting did not contain the scripts for the listening passages, so a separate form was created for the listening test which contained both the listening scripts and the associated items. In the training session, judges took the test using the training form. During the operational standard setting of the listening test, judges were not able to hear the taped version of the questions but were also provided with the scripts of the listening questions. The week prior to the training session, an email was sent to all judges that contained

1. an introductory letter with a link to a CEFR familiarization website, www.CEFtrain.net
2. an agenda for the training session, consisting of adapted versions of pages 33-36 from the CEFR (2009).
3. the training materials, the listening and reading components of the CEFR (2009) self-assessment grid (CEFR Table 2); and a link to the website.
4. two forms collecting personal information and agreements concerning test security and informed consent for the research portion of the project. (see Appendix 2 and 3)

As homework, judges were asked to refer to the website and level summaries, and use the self-assessment grid to assess themselves (in any second language) and their students, in terms of the CEFR levels. (Council of Europe, 2009).

Training of judges was extremely conventional and followed suggestions given in such authoritative sources as Cizek (2001), Cizek and Bunch (2007), and the Council of Europe (2009).

Day 1 of training-Introduction to the CEFR

On the first day of training, judges were given a brief PowerPoint presentation explaining the purpose of the project, a description of the EPT and an explanation of how it was developed and validated. Following guidelines provided by Cizek (2001), Cizek and Bunch (2007) and the Council of Europe (2009) a great deal of effort was extended during training to familiarize judges with the descriptors used for the panels. They then took part in a CEFR familiarization process. After a brief description of the CEFR, their understanding of descriptors was tested. Judges were given a sheet containing the Global Descriptors from the CEFR Table of Global Descriptors. The descriptors had been rearranged, and the judges were asked to sort them back into the correct order (first individually, then in pairs). After providing them with a copy of the original CEFR Table and discussing the correct answers, the judges were asked to take out their 'homework' activity in which they rated their own ability and that of their students using the CEFR levels, and to discuss their answers in pairs.

PLD Test of CEFR Descriptors

The session then shifted to the CEFR reading Performance Level Descriptors (PLDs). The first activity was another sorting activity, in which judges were asked to sort 20 CEFR reading and 20 CEFR listening descriptors from CEFR levels A1 to B2. They were then given a sheet containing CEFR reading descriptors from the scales used in the study, for CEFR levels A1 to B2. These descriptors were in a randomized order. Judges were asked to sort the descriptors into an order from least difficult to most difficult that they felt made the most intuitive sense and assign a CEFR competency level to each descriptor.

The training for the listening PLDs was conducted in parallel fashion. Judges were asked, individually to sort 20 CEFR listening descriptors taken from the CEFR A1 to B2 levels. After they finished, correct answers were provided along with a full list of the listening PLDs from the scales used in the study. The scores on these activities were recorded and analyzed later as a measure of how well the judge could use the CEFR descriptors for levels A1 to B2.

Discussing Difficulty

After a break for lunch, judges took the practice test that was described above. The judges were then divided into the three groups of six people in each of the operational panels. The judges were asked to sit together in a circle with the other members of their standard setting panel. A group leader was chosen, and each panel was asked to go through the test form, item by item. As a group they were asked to discuss what knowledge, skills and abilities were required to answer each item, and how the items differed in terms of difficulty. One hour and fifteen minutes was allotted to this task. The discussions were taped by the facilitators for later transcription.

Discussing the Barely Proficient Student

Following this activity, the judges were introduced to the concept of the barely proficient B1 student (B1 BPS). They were then given a form which contained space for their notes on the BPS and told to refer to their listening and reading PLDs for the A2 and B1 levels, and summarize on the forms what they felt to be the key characteristics of a B1 BPS for both listening and reading. They were then asked to discuss their summaries in pairs or small groups. This was the final training activity of the day. Judges were then given the opportunity to ask any questions they had about what had been discussed to that point. They were told that when they returned for the

actual meeting, they would have one training round prior to the meeting, then they would perform the actual standard setting. This concluded the Day 1 of the training session.

Day 2 – The Operational Standard Setting and Review

The Angoff meetings were held over the period of one week on July 12, 14 and 16 in 2010. The meetings were divided into two panels with standards set for the reading test in the morning and the listening test in the afternoon. Before beginning, judges were given a brief review of the contents of the previous training session. This included a review of the B1 BPS. Judges were then told to estimate, based on their understanding of students in the PE program (or Taiwanese university students in general for judges who were instructors at other universities), the percentage of students who had reached the B1 level for the skill in question and write down this estimate. The test form and the Round 1 rating form (see Appendix 4) for the reading test were distributed to the judges and a practice round was conducted.

Day 2 – Round 1

The rating form contained a single column for each item being rated with each column containing a list of probabilities in increments of 0.1, starting from 0.1 to 0.9 with a space between each figure. Judges were asked to "circle or insert" the probability that a just-B1 level student would answer the item correctly, and to write their answer at the bottom of the column (see Appendix 4).

Judges were instructed not to attempt to include guessing in their calculation of probabilities. They were then given a practice round, in which they were asked to write their ratings for the first few items. It was made clear this was simply a practice round, to ensure that they understood the procedure and that they could change their answers later. The facilitators

circulated from judge to judge while they were performing the practice round to make sure the procedures were understood. Once all judges had finished, they were asked if there were any remaining questions. After questions were answered, the first round of ratings was then conducted.

After returning from a break, judges were given forms containing both feedback data and empirical item-difficulty data. They were given feedback data in the form of a distribution of actual students in the program at different scores levels on the test. The rating form for the second and third rounds incorporated further feedback. The range of scale scores was divided into 40 categories of approximately equal size. A column was added to the left side of the form. Each row in the column contained one of the 40 categories, from low to high. Once again, there was one column per item and the columns contained probabilities in increments of 0.1. This time the probabilities were placed in rows corresponding to the scale scores in the left most columns. Based on empirical results from the spring 2010 administration of the EPT, the probabilities were placed in the particular scale-score row to correspond to the approximate probability that a student in that scale-score category would answer the item correctly. Judges were guided in the use of the feedback material, so that they could use their initial estimates of students at the B1 level, the distributional data and the second rating form to contrast their Round 1 rating with what their rating would have been based on their estimate of the number of students at the B1 level. Finally, at the bottom of the column for each item was the empirical p-value for all PE students who took the midterm EPT. The listening form also contained this information for graduating students. For reading, the difference between graduating students and all students was not large, so this information was omitted.

After being instructed in the use of the feedback information, a discussion session was held. For each item, the judges announced their Round 1 ratings and briefly explained why they had given the rating to each of their items. The assistant facilitator entered ratings into the computer as they were announced. Once the discussion period was finished, the cutscores were calculated and shown to the judges. Using the distribution data, judges were asked to contrast the percentage of students they had initially estimated to be at the B1 level with the percentage of students who would be classified at the B1 level based on their round one rating. They were then asked to make their Round 2 ratings. It was emphasized that they did not need to change their ratings.

Day 2 – Round 2 and 3

The Round 2 ratings were entered into the computer and cutscores were calculated. (There was no discussion of individual decisions following Round 2; rather, judges handed their rating forms to the facilitators who entered their ratings into the computer while those who had finished took a break.) Judges were again asked to consider the impact (distributional) data, and given the opportunity to ask questions or make comments. Following this, they were asked to make their final round of ratings. The ratings for the final round were used to derive the recommended cutscores.

At the opening training meeting, all participants were asked to sign a research consent form releasing all the data generated from the standard setting to the school for any research and administrative purposes that were necessary (see Appendix 2 and 3). In addition, judge's feedback about their familiarity and confidence and understanding of the training was gathered regularly throughout both training and the operational standard setting panels.

Summary of Day 1 and Day 2

Day 1

1. pre-training assessment of individual preparation (Appendix 5)

2. 3 different assessments throughout the training day assessing confidence in and familiarity with their task (Appendix 6, 7, and 8)

Day 2

3. An assessment at the opening of the operational panel to address confidence and preparation in the day's coming activities. (Appendix 9)

4. Three rounds of the operational standard setting, reading panel

5. Lunch

6. Three rounds of the operational standard setting, listening panel

7. A final assessment (Appendix 10) of judge's confidence in their final cutscore decision and satisfaction with the manner in which the standard setting training and panels had been conducted. This was modeled after the sample form contained in Cizek & Bunch (2007).

Focus Group and Recording

In addition to the feedback forms, the group discussion activities described earlier were recorded and later transcribed. Following the operational standard setting, Group 3 volunteered to take part in a focus group to discuss their impressions of the standard setting. This focus group was recorded and later transcribed for use in understanding judge's perceptions of the standard setting, its procedures and its outcomes. These recordings were made by hand-held analog tape recorders with full knowledge of all the participants. Full disclosure of all data gathering practices was conducted throughout.

3.4 Assessment Tools and Measurements

3.4.1 Assessment Tools

The three main measurement procedures used to assess judge accuracy were the p-value correlation, the Root Mean Square Error (RMSE), and Mean-Number Difference (MND).

p-value Correlation

P-value correlation refers to the correlation of the item estimates made by the judges and the actual p-value of the item measured under high-stakes conditions. In a widely cited review of the Angoff method, Brandon (2004) describes the p-value correlation or, as he labels it, "deviations of judges from empirical p-values" (p. 72). He cites this in a section of the paper dealing with studies of the validity for the Angoff method, Drawing from Kane's (1994) theory of standard setting validity, Brandon states that Kane's 'procedural validity' can be assessed through examination of item estimates, one of which includes "correlations of the set of mean item estimates with the set of p-values before any review of empirical information or discussion among judges had occurred": in effect, Round 1 (p. 71). Hambleton (2001, p. 112) cautions about the use of this measure in the later rounds of the standard setting, "Often a very high correlation between judge's ratings and candidate score information is taken as evidence that the empirical data are driving the standard setting process." However, Brandon, (2004, p. 72) goes on to state the more common conclusion that, "The smaller the deviation, the stronger the evidence for validity".

RMSE

The Root Mean Square Error or RMSE is an estimate of the standard error of the mean that can be used in a number of measurement situations including Angoff method estimates across items and judges. It represents the variation in the cutscore when the same items are judged by a different sample of judges (Verhoeven, et al., 2002), and indicates the error involved in the test's passing score (Schoonheim-Klein et al., 2009). The formula for the RMSE is shown below.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \left(P_i - O_i\right)^2}{n}}$$

Where

P_i = the p-value predicted for the item by the judge

O_i = the empirical p-value value of the item as observed during high stakes use.

n = the total number of cases in the sample being examined

The p-value correlation, like all measures based in correlation, is a measure of the relative relationship between the two variables in question. In comparison, the RMSE, is a measure of the absolute relationship between the variables. Use of both values allows for a more complete comparison of variables.

The RMSE is not widely used in standard setting. For example, a Google scholar search of "RMSE + Angoff" found only 59 results, none of which were leading papers in the field. Nevertheless, the RMSE is used in other fields such as medicine and biology, as well as fields

that require forecasting. While there are a large number of alternative measures available, the RMSE is the most widely used and best understood (Hyndmana & Koehlerb, 2006)

Cut-Off Score Judgment (CSJ)

The Cut-Off Score Judgment or CSJ was developed as a measure of the average distance from the grand mean of all the estimates from the individual estimate made by the judge on each of the individual items. Similar to the RMSE, it is an absolute measure. It differs from the RMSE in that the RMSE squares the difference between the measured p-value and the estimated p-value. As a result, direction of the error is not important for the RMSE. It is a measure of the absolute value of the difference between the grand mean of the each round and the individual estimate made by each of the judges for each item. The CSJ maintains direction in it measurement. The value of the difference between the measured and the predicted p-values is not squared, and as a result, can be either positive or negative.

Like the RMSE, the CSJ is not widely used in standard setting research.

3.4.2 Assessment Procedures

Prior to the beginning of the standard setting, judges were mailed a Panelist Information Form. A copy of this form can be found in Appendix 5. Most of the information requested on the form was demographic, although some of the questions refer to knowledge of testing.

In addition to the items selected from the operational EPT and the mock EPT used during Day 1 of training, this study used four different measurement scales and activities. These included

1. a test of matching CEFR items with statements about linguistic competency

2. assessment of knowledge of standard setting and the Angoff method
3. assessment of knowledge of the CEFR
4. assessment of knowledge of the Practical English Test

Day 2 used a further two evaluations of the validity of the Angoff method and the relationship with training. These included,

5. an evaluation and overview of the procedure at the beginning of Day 2
6. a final evaluation of the activities of the standard setting procedure at the end of Day 2

ASSESSMENT 1, Day 1 Knowledge of CEFR Descriptors

The first measurement taken during the study was the PLD sorting procedure described above. Judges were given a sheet of paper that listed 20 CEFR reading descriptors and another sheet that listed 20 CEFR listening descriptions taken from the CEFR listening and reading Global Descriptors for CEFR levels A1 to B2. They were then asked to write down the CEFR level of the 40 CEFR Global Reading and CEFR Global Listening Descriptors into the order that the CEFR would rank them.

Cronbach's Alpha was calculated for the reading descriptor scale and for the listening descriptor scale. Usually, in this sort of study, Alpha is set at 0.70. However, the items were descriptors taken from the handbooks of the CEFR handbook. (Council of Europe, 2001, 2009). The handbooks state that the "Can do" statements of the CEFR standard are not intended to be used as a proficiency scale. Rather, they are intended for use as a statement of learner competence. As a result, items may not be psychometrically sound. In fact, the listening test was unable to reach Alpha = 0.70 and the reading test could only obtain Alpha = 0.70 by deleting a large number of items. As such, the criteria for inclusion to the scale were set at 0.60.

The Alpha for the 20-item listening scale was 0.49. Two items was deleted one-at-a-time according to the suggestions provided by SPSS20. The final Alpha for the 18-item scale was 0.61. Similarly, the Alpha calculated for the 20-item reading scale was -0.12. Seven items were deleted one-at-a-time according to the suggestions provided by SPSS20. The final Alpha for the 13-item scale was 0.71. Scores for Alpha reported in this study are for the 18-item listening scale and for the 13-item reading scale.

ASSESSMENT 2, Day 1 Confidence in Standard Setting Procedures

The second measurement was an 8-item scale constructed solely for this study, although items resemble those found in Cizek (2001). Cizek's later book (2012) supplies a wider range of item possibilities, but these were not available when this study was done. Loomis (2012) reports that NAEP has a series of questions that it has been asking since their standard setting studies began. Details about the results of these questions, either by Cizek or Loomis, are not reported at all, so there is no way of comparing their results with those of this study. Likert-scale scores selected by the judges were used to calculate judge's confidence in their knowledge of standard setting procedures and the Angoff method in particular.

Assessment 2 is aimed at measuring judge's basic knowledge about standard setting after being introduced to the idea and to the Angoff method of conducting the procedure. A copy of the form can be found in Appendix 6. With an N=18, the Cronbach's Alpha for the assessment is 0.70.

ASSESSMENT 3, Day 1 Familiarity with the CEFR

Assessment 3 is a 9-item scale constructed solely for this study and designed to assess judge's basic familiarity and opinions following training about the CEFR. The form can be found in

Appendix 7. It should be noted that this was not designed to assess judge's ability to use the CEFR properly. Rather it was designed to measure judge's confidence in their understanding and ability to use the CEFR. With an N=18, the Cronbach's Alpha for this scale is 0.81.

ASSESSMENT 4, Day 1 Familiarity with the Practical English Program

Assessment 4 is an 8-item survey designed to provide the standard setting facilitators with an overview of how well the judges believed they understood the events of the entire Day 1 training. The items of Assessment 4 are contained in Appendix 8. The Cronbach's Alpha for Assessment 4 is 0.79.

ASSESSMENT 5, Day 2 Morning Confidence in Training Skills

Assessment 5 was the first activity conducted on the second day of the standard setting and was conducted as part of a warm-up session to remind judges about the procedures. The main research purpose of the activity was to survey judge's feeling of confidence as they entered the operational part of the standard setting. Assessment 5 was an 8-item survey of judge's perceptions of their current confidence with the procedures and skills that had been covered on the previous day's training (see Appendix 9). The survey was not intended as a scale, although all of the questions were connected to an understanding of the various parts of the Angoff procedure and the instructions the day before. Judges were asked if they felt qualified to conduct the standard setting and in another question, were asked if they felt they were ready to start the procedure. Likert-scale scores selected by the judges were used to calculate judge's confidence of their knowledge of standard setting procedures and the Angoff method in particular and calculate judge's confidence in their preparation to perform a standard setting.

Even though the questions were not intended as a scale, treating it as a scale and calculating the Cronbach's Alpha for the scale produced an Alpha = 0.90.

ASSESSMENT 6, Day 2 Final Evaluation of Confidence in Overall Performance

Assessment 6 was the measure of judge's overall perceptions of the standard setting. More than any other assessment, this was not a scale (see Appendix 10). Likert-scale scores selected by the judges were used to calculate judge's confidence in their performance during the standard setting.

Once again, even though the questions in the evaluation were not designed with the idea of being a scale, if we treat Appendix 10 as a scale, the Alpha – at 0.86 – was high.

3.5 Assessment Expectations

The goal of this study was to determine the efficacy of training on the ability of judges during an Angoff standard setting. Raw scores were treated as the unit of analysis and not converted using latent trait models, such as Rasch modeling, etc. Two aspects of training were investigated to see if they had an effect on the final performance of the judges; the efficacy of the training and the self-reported familiarity of judges.

3.5.1. Efficacy

Effectiveness of performance was determined through a series of different types of assessments. The first of these was Assessment 1. Although this was done as part of the training exercises, it was intended as a measure of judge's understanding of the PLDs.

Efficacy was measured using PLDs through comparison of p-value correlations and Root Mean Square Error (RMSE). P-value correlations were determined by calculating the Pearson correlation between the empirical p-value of the item measured during actual use of the item during high stakes tests and the judge's estimate of the p-value obtained during the operational standard setting (Brandon, 2004). The RMSE was calculated for the performance for each member of a panel.

3.5.2. Confidence & Familiarity

The degree of confidence and familiarity held by individual judges was measured using a series of self-report surveys. This included Assessment 2, 3, 4, 6 and 7 (Appendix 6, 7, 8, 9 and 10). These were used in the fashion recommended by leading researchers in the standard setting field, such as Cizek (2012), Hambleton et al. (2012) and Loomis (2012). Raw cutscores were created

using the Angoff method of standard setting from the performance of the judges. These raw scores were compared directly using correlations and ANOVA. The results of significance tests were reported where applicable

3.6 Data Analysis

Data analysis was performed with several statistical tests that are described below.

3.6.1 Repeated Measures Analysis of Variance

Repeated measures analysis of variance (rANOVA) was performed initially on the data to determine if the feedback portion of the standard setting between rounds had the intended effect of altering judge's item estimates. The rANOVA is performed separately on each of the three different measures of judge accuracy to assure that judge's scores responded to the announcement of feedback data between the different rounds in the direction suggested by the Angoff model. It is expected that if the three different measures of judge accuracy had been affected as intended by the feedback, the within-subjects measures of the rANOVA would show a significant difference between the rounds where the feedback data was announced. In addition the power of each of the rANOVA is reported.

3.6.2 Pearson Correlation

Pearson correlation were performed in a series of analysis to test the main research hypothesis of this study, as well as illustrating the convergent validity of the different measures used in this study. Significance testing is used in each case based on the n size of the analysis to determine if the correlations are significant at the .05 level.

.

CHAPTER 4 RESULTS

In the following tests, significance is assumed to be $p < .05$. This value was chosen because of the small number of judges (N=18) used in the standard setting. Selecting the higher value of $p < .01$ makes significance much harder to obtain, and hence values that could indicate meaningful difference might become lost. However, choosing a significance level of $p < .05$ requires careful reflection. Five percent of correlations will reach significance simply by chance. In addition to the correlations reported here, a large number of exploratory correlations were performed. With this in mind, patterns of findings need to be interpreted with caution. All statistics and numbers are rounded to the second decimal place, where appropriate. Cutscore statistics for the standard setting are contained in Appendix 11 and 12.

The Performance of Different Angoff Panels

Table 4.1 shows the p-value correlation, the Root Mean Square Error (RMSE), and the Cutscore Judgement (CSJ) for each of the judges on the reading panel. Table 4.2 shows the same values for the listening panel. A one-way repeated measures ANOVA was performed separately for the reading and listening tests for each of the measures of judge's accuracy (ie: separate analysis of the p-value correlations, the RMSE, and the CSJ). The purpose of this analysis is to determine if the feedback data shown to the judges between rounds had the effect that was intended. As such, it is expected that the within-subjects measures for each of the three separate measures of judge accuracy would show a significant difference between the three different rounds of the standard setting.

Within-subject effects were significant for all of the measures of judge accuracy for the reading panels (Table 4.1). For the p-value correlation measure of the reading panel. Wilks'

Lambda = .162, F(2,16) = 41.50, p < .01, η^2 = .84. For the RMSE, Wilks' Lambda = .373, F(2,16) = 13.47, p <.01, η^2 = .627. For the CSJ, Wilks' Lambda = .620, F(2,16) = 4.91, p = .02, η^2 = .38.

Within-subject effects were significant for all of the measures of judge accuracy for the listening panels (Table 4.2), except the CSJ. For the p-value correlations of the listening panel, Wilks' Lambda = .17, F(2,16) = 39.40, p < .01, η^2 = .83. For the RMSE, Wilks' Lambda = .32, F(2,16) = 17.40, p < .01, η^2 = .69. For the CSJ, Wilks' Lambda = .85, F(2,16) = 1.43, p = .29, η^2 = .15.

All of the values of the within-subjects ANOVA results were significant, except for the CSJ listening measures in Table 4.2, which was not significant. Some caution should be taken in interpreting the individual results for each of the subjects in Table 4.1 and 4.2. Particularly in Table 4.2, which is the Listening Panel, variation in the measures is very large. As a result, the ANOVA is not significant. However the standard deviations of the between-subject grand means, as expected, decreases between Round 1 and Round 2 for both the reading and the listening panels. And while the standard deviation does not decrease for Round 3 of the reading panel, the value is essentially the same as the value obtained for Round 2.

This indicates that generally the standard setting feedback between rounds 1 and 2 and between rounds 2 and 3 had the intended effect. This point will be returned to in the conclusion, although it points to some problems with the CSJ measures of the listening panel and that these should be interpreted with some caution.

Table 4.1

Measures of Judge Accuracy - Reading

	p-corr R1	p-corr R2	p-corr R3	RMSE R1	RMSE R2	RMSE R3	CSJ R1	CSJ R2	CSJ R3	PLD Test Score
Panel 1										
Agf11	.33	.77	.79	28.77	16.62	16.79	12.53	1.60	2.50	3
Agf12	.54	.90	.97	19.32	14.44	11.29	-1.50	3.33	2.05	7
Agf13	.57	.58	.75	19.31	20.02	19.09	4.25	1.50	5.13	8
Agf14	.37	.66	.66	23.11	21.14	21.14	4.00	6.25	6.25	5
Agf15	.43	.87	.92	19.83	12.26	13.91	-3.13	-2.63	2.63	7
Agf16	.21	.45	.68	23.85	25.11	23.05	-0.38	8.88	7.63	3
Panel 2										
Agf21	-.04	.27	.16	28.78	23.78	25.38	-18.8	-15.7	-16.0	12
Agf22	.31	.41	.41	26.12	24.06	25.33	8.60	6.60	8.72	8
Agf23	.43	.79	.76	20.33	16.73	18.33	-1.75	2.00	3.50	10
Agf24	.06	.07	.48	23.08	22.99	19.00	-2.33	-2.38	-5.38	4
Agf25	.12	.44	.42	31.64	22.05	21.61	-28.2	2.63	-3.88	10
Agf26	.30	.37	.47	22.05	21.50	20.46	-1.13	-0.38	0.25	6
Panel 3										
Agf31	.63	.61	.71	19.83	20.27	18.86	2.88	3.25	2.63	8
Agf32	.54	.59	.87	20.33	19.89	15.33	0.75	1.75	3.25	8
Agf33	.41	.86	.94	22.23	12.55	13.35	-9.25	-1.63	2.38	4
Agf34	.55	.66	.76	19.94	17.97	16.71	-3.50	0.20	1.83	7
Agf35	.38	.85	.95	21.57	16.57	14.14	-9.93	3.68	3.93	4
Agf36	.65	.92	.97	17.32	14.18	15.10	-8.00	3.38	5.47	4
Grand Mean	.37	.62	.70	22.63	19.01	18.27	-3.05	1.24	1.83	6.56
SD	0.20	0.24	0.23	3.88	4.02	4.07	9.51	5.21	5.64	

Table 4.2
Measures of Judge Accuracy – Listening

	p-corr R1	p-corr R2	p-corr R3	RMSE R1	RMSE R2	RMSE R3	CSJ R1	CSJ R2	CSJ R3	PLD Test Score
Panel 1										
Agf11	.45	.79	.87	20.39	15.16	13.49	2.61	0.73	0.36	6
Agf12	.72	.93	.98	17.45	11.32	7.78	2.24	-1.14	-2.52	11
Agf13	.43	.82	.87	24.40	17.93	15.13	8.76	4.91	2.79	14
Agf14	.56	.86	.85	18.51	14.25	16.48	0.73	1.36	3.99	8
Agf15	.37	.72	.72	21.01	17.41	17.40	2.86	2.36	2.24	13
Agf16	.47	.61	.91	18.69	18.49	18.44	0.11	1.86	6.11	12
Panel 2										
Agf21	.28	.67	.67	18.31	13.87	13.91	-9.52	-7.64	-8.02	14
Agf22	.42	.50	.68	21.98	20.96	19.38	5.11	4.49	4.49	10
Agf23	.55	.83	.71	16.05	15.87	10.30	-4.64	2.11	10.36	8
Agf24	.53	.71	.95	21.12	16.38	5.89	-22.7	-18.5	-7.64	5
Agf25	-.06	.53	.42	23.22	18.33	20.68	-2.22	-13.0	-10.1	9
Agf26	.69	.69	.77	15.11	15.11	13.08	-2.52	-2.52	-4.02	7
Panel 3										
Agf31	.66	.67	.70	19.58	17.32	17.32	-0.27	0.23	0.48	10
Agf32	.56	.78	.78	15.96	13.47	13.47	3.11	3.49	3.49	5
Agf33	.65	.93	.95	19.35	17.79	16.46	-1.52	1.49	0.98	6
Agf34	.60	.69	.74	19.52	15.71	16.36	3.36	2.74	1.74	6
Agf35	.39	.94	.95	23.35	16.62	16.64	0.23	3.96	4.91	6
Agf36	.46	.93	.93	16.83	17.02	17.02	7.49	5.24	5.11	13
Grand Mean	.49	.75	.79	19.49	16.28	14.91	-0.38	-0.44	0.82	9.05
SD	0.18	0.14	0.14	2.69	2.23	3.87	7.03	6.41	5.40	

Issues Concerning the p-value Correlation

It is important to note that some of the correlations in Table 4.1 and Table 4.2 during Round 3 are as high, or even higher, than .90. Such high correlations for this kind of task are improbable and are seldom achieved under conditions where participants do not have the kind of feedback information and discussion typically provided for standard setting judges (Brandon, 2004; Clauser et al., 2002; Clauser et al., 2009; Clauser et al, 2013; Clauser et al., 2014; Margolis & Clauser 2014). There are two explanations that have been suggested for such high correlations.

The first of these can be thought of as the anti-high correlation position. The anti-correlation position is derived from experimental results. In fact, this position is supported by data from a large number of related studies that indicate judges are typically unable to accurately estimate the difficulty measures of items when asked to do so (Goodwin, 1999; Impara & Plake, 1998; Lorge & Krulou, 1953; Linn & Shepard, 1997; Norcini et al., 1987; Norcini, Shea & Kanya, 1988; Shepard, 1994; Smith & Smith 1988; Taube, 1997).

Brian Clasuer and his colleagues have tried to address this problem experimentally. For example, in one study (Clauser, et al., 2009) judges were shown items and asked to rate their difficulty values, then asked to discuss their estimates with other judges without the benefit of feedback information. After decisions were made without feedback, judges were then provided with feedback information similar to that provided in an operational Angoff standard setting (Clauser, et al., 2009, p.1). Discussion of the items decreased the variance of the judges' estimates, just as it would during a typical standard setting, but it did not improve the relationship between the judgments and the difficulty values of the items. However, once the judges were given performance data for a subset of the items, judgments of the items showed a substantial increase in their correspondence with the true difficulty measures of the items. Clauser has interpreted

this kind of result as demonstrating that without performance data, judges are unable to accurately gauge the difficulty of items. As such, judges in an operational Angoff standard setting will not be able to accurately estimate the difficulty of an item without first being told how well students perform on the item.

The pro-high correlation position is derived from Item Response Theory (IRT) and the IRT formula. It proposes that because judges are not able to accurately calibrate the difficulty scores for the items, they need additional information to calculate the minimum value of the cut score for the item. High p-value correlations are thus a necessary part of a correctly conducted Angoff standard setting, demonstrating that judges understand their role as judges, as well as the role of feedback information (Cizek, 2012; Hambleton et al., 2012; Loomis, 2012).

This position is consistent with the interpretation of the Angoff method from a psychometric point of view.

If $p_i = \exp(\Theta_j - b_i) / 1+\exp(\Theta_j - b_i)$

Where,

p_i = the estimated probability of a correct answer

Θ_j = the judge's true mental image of the ability level for the borderline candidate, in this case, the CEFR B1 level

b_i = the true difficulty of the item under high stakes conditions

Feedback information appears to be a necessary but not sufficient condition for an accurate judgment. Providing the feedback information merely allows the judge a greater understanding of b_i. Since the real value of Θ_j remains unknown, knowing the value of b_i merely gives the judge greater ability to estimate the range of true difficulty values for Θ_j, but does not tell them what it would be. This is particularly important for teachers whose students are drawn from special populations of high or low performers and whose estimates of Θ_j might be skewed as a result.

As mentioned earlier, a focus group was conducted with Group 3, where the issue of feedback data was raised by the moderator. Data from this focus group support the idea that judges use the feedback data as a source of additional information to make decisions about their estimates of student ability.

Moderator: *One of the things I'm particularly interested in is your impression of the impact data [feedback data]…What you thought of that when I presented it. How that affected you when you made your decisions about things…*

Judge Agf 36: *For the morning part…it shows that I am underestimating the students…I underestimate them, their ability. I underestimate their test taking skills…*

Moderator: *So you saw it as a source of additional information?*

Judge Agf 36: *Yes*

Moderator: *So how did you use that in reevaluating their scores…Did it have any impact on your image of the B1 student?*

Judge Agf 36: *Yes…I think it increased the percentage a little bit. Not just the original student in my mind. I started to think about, not just the original PE midterms and finals, those who can get more than 80. In the beginning, I started to think about 90…*

Moderator: *Did it have any impact on the idea of what a B1 student…*

Judge Agf 36: *I was thinking the student will make this or that mistake. And when I compare to general student performance, I kind of find out my score kind of matched the PE more than the B1 level. So then even though I was doing it unconsciously, then I feel maybe I was thinking about the average student instead of the B1 student. So then when I do the listening, I kind of raise it [the estimate of B1 cutscore]*.

Moderator speaks now to Judge Agf 31.

Moderator: *Now you teach at another school, and your students are different from ours in many ways. Did the impact data help you to understand the relationship between the test and the students better?*

Judge Agf 31: *Yes.*

Moderator: *In the way that Judge Agf 36 was talking about?*

Judge Agf 31*: Yes…When I saw the results you provided for us that helped me to understand the real performance for our students. That would be maybe higher than the B1 according to the descriptors, so I raised my criteria for the second level.*

Despite theoretical and empirical reasons to accept the value of feedback, there is still one possibility that Clauser and his colleagues may be bringing to the interpretation of the standard setting. It still remains possible that tasks involving the incorporation of feedback data into the mental calculations demanded of standard setting judges are so mentally difficult, that at least some judges are not able to perform the task without losing track of the performance standard and it descriptors. So rather than use the feedback data to mentally position their students relative to the cutscore, they don't use the performance standard at all and just follow the ups and downs of the feedback data. This, however, was not what the judges in Panel 3 described doing.

1. Does knowledge and training in Performance Level Descriptors (PLDs) work effectively to predict an Angoff standard setting judge's ability?

Performance Level Descriptors (PLDs) are among the key concepts in describing the mental work of a standard setting judge. In an Angoff standard setting, judges must match real items with their mental image of a description of a person at the cutscore. In the IRT model of the p-value formula shown above, this ability to identify the items that represent the judge's mental image of a person who is barely at the cutscore are represented by Θ_j. Thus, for a judge to perform accurately, they must have a clear and consistent image of Θ_j. The assumption that training in the PLDs of the standard setting produce better judges is implicit in the design of the Angoff standard setting method (Council of Europe, 2009). Large portions of the Council of Europe 2009 manual are devoted to the theory and practice that knowledge about PLDs will improve the judge's ability to perform their responsibilities in the standard setting (Council of Europe, 2009). Large portions of the training for this standard setting were committed to this assumption. Certainly the idea carries with it a great deal of face validity. Is it supported by empirical findings?

The 18 judges were separated into 3 panels of 6 people each for the purpose of training and tested on their ability to match levels of the Common European Framework of Reference with descriptors taken from the CEFR Scales. This assessment is described in greater detail in Assessment 1. Tables 4.1 and 4.2 show the number of descriptors correctly categorized for each of the judges according to their group membership.

As Table 4.1 illustrates, the best performing panel in terms of correctly identifying CEFR reading descriptors is clearly Panel 2. Despite this, by Round 3, the Group Mean for their p-value

correlation (r = 0.45) was the lowest of all three groups (Panel 1, r = 0.80; Panel 2, r = 0.87), and their RMSE remained higher (RMSE = 21.68) than the other two groups (Panel 1, RMSE = 17.55; Panel 3, RMSE = 15.58). These results indicate that knowledge of the PLDs did not assure that a judge can accurately estimate the difficulty of an item.

The scores from Table 4.2, the listening descriptors, were slightly different. While Panel 2 continued to measure lower than other panels on the p-value correlations (r = .40, .66, and .70), Panels 1 and 3 showed a different pattern (Panel 1, r = .50, .79, and .87; Panel 3, r = .55, .82, .84). This was especially true for the RMSE. While the Group Mean for Panel 1 on the listening panel started out slightly higher (RMSE = 20.10) than the Group Mean for Panel 3 (RMSE = 19.10), by the end of Round 3, Panel 1 had a slightly lower Group Mean (Panel 1, RMSE = 14.79; Panel 3, RMSE = 16.07), but in fact, their performance was about the same during each of the rounds.

Table 4.3 shows the Pearson correlation between the PLD‘s Test scores and accuracy of the judge’s during the standard setting. For an $N = 18$ ($r = .44$, $p < .05$). Although there are some singularly high correlations, the absolute value of only 3 of the 18 correlations reaches significance. This shows that the relationship between knowledge of the PLDs and the three measures of judge accuracy is not strong, and that there may be other important variables that are unmeasured in this model. In conclusion, the PLD Test is not an important predictor in determining judge accuracy.

Table 4.3

Correlation between PLD Test and Standard Setting Performance

	p-corr	p-corr	p-corr	RMSE	RMSE	RMSE	CSJ	CSJ	CSJ
Round	R1	R2	R3	R1	R2	R3	R1	R2	R3
READING	-.14	-.23	-.48*	.20	.23	.36	-.38	-.45*	-.48*
LISTENING	-.29	-.13	-.13	.12	.18	.28	.32	.17	.02

**indicates p<.05*

These results seem to suggest that there is little or no relationship between a judge's knowledge of how to use the PLDs and their performance in an Angoff standard setting. As such, even if a judge performed poorly on a test of PLDs, they would still be able to act competently as a judge in such a standard setting. This is easier to understand if you consider that a standard setting involves a large number of skills other than just knowledge of the PLDs. Lack of understanding of the PLDs could be compensated for by strengths in other areas of performance. On the other hand, knowledge of the PLDs could be a necessary but not sufficient condition for competent performance as an Angoff standard setting judge.

A second explanation for the poor relationship between the PLD Test and judge's standard setting performance addresses the test itself. A great deal of difficulty was observed in the construction of the PLD Test. A number of the original items had to be deleted to create a psychometrically acceptable measurement instrument. It should be noted that a formal test was never developed and the PLD Test was used only because its Cronbach's Alpha was acceptable. Many indexes of reliability and performance were not checked. It is possible that the items used to test the judge's PLD knowledge do not provide a completely sound measurement instrument.

Table 4.4 shows the correlation matrix for the standard setting that examined reading ability. The matrix provides convergent validity (AERA, APA, & NCME, 2014; Cronbach, 1988; Cronbach & Meehl, 1955; Kane, 2006; Loevinger, 1957; Messick, 1981, 1989, 1998) for the results of this standard setting. Correlations of RMSE scores with other RMSE scores should theoretically be positive. The same would be true for p-value correlations with other p-value correlations. On the other hand, all correlations of p-value correlation scores and RMSE scores should theoretically be negative. Correlations with the CSJ could either be positive or negative for either the p-value correlation or the RMSE. All of the measures of judge accuracy between the RMSE and the p-value correlation were significant. While measures of CSJ were highly correlated with each other at different rounds, they were only moderately correlated with p-value correlation, with 5 of 9 correlations reaching significance. With RMSE, only 1 of the 9 correlations was significant.

In Table 4.4, the absolute value of correlation scores range from $r = 0.48$ to $r = 0.89$. By Round 3, the end of the reading panel, the absolute value of 5 of the 8 correlations exceeded significance at $r = 0.44$. The only correlations that failed to meet significance were those between the CSJ and RMSE.

In addition, all the correlations between p-value correlation scores and RMSE scores were negative and correlations between the same measures of ability were positive.

Table 4.4

Matrix Correlation for Measures of Reading Ability

	p-corr 1	p-corr 2	p-corr 3	RMSE 1	MSE 2	RMSE 3	CSJ1	CSJ2	CSJ3
p-corr 1	1.00	.71*	.77*	-.79*	-.58*	-.61*	.41*	.43*	.65*
p-corr 2	X	1.00	.87*	-.48*	-.88*	-.75*	.11	.28	.50*
p-corr 3	X	X	1.00	-.68*	-.82*	-.89*	.26	.41*	.60*
RMSE 1	X	X	X	1.00	.50*	.60*	-.32	-.22	-.46*
RMSE 2	X	X	X	X	1.00	.89*	.00	.03	-.23
RMSE 3	X	X	X	X	X	1.00	-.05	-.09	-.24
CSJ 1	X	X	X	X	X	X	1.00	.43*	.61*
CSJ 2	X	X	X	X	X	X	X	1.00	.88*
CSJ 3	X	X	X	X	X	X	X	X	1.00

**indicates p<.05*

Table 4.5 shows the same matrix of correlations for judge accuracy of listening ability. The correlations in this matrix are not as large as those in the judge's reading ability matrix. The absolute value of the correlation scores range from r = .22 to r = .84, although the highest correlations come from measures of CSJ with measures of CSJ in other rounds. Only 13 of the 36 correlations reach significance at r = .44, three of those correlations came from measures of CSJ correlating with measures of CSJ in other rounds. Measures of CSJ all correlated significantly with other measures of CSJ during other rounds, however only 1 of the correlations with a measure of RMSE was significant and none of the correlations with p-value correlation were significant.

As in the judge's listening ability matrix, all the correlation scores are in the expected direction with correlations between p-value correlations scores and RMSE scores all negative, and correlation scores between the same type of measure of judge ability all positive. However, at Round 3, only 3 of the final 5 correlations are significant.

The data indicates that the p-value correlation and the RMSE are not measuring the same things. If the correlations were very high, approaching unity, this would indicate that the two measures were probably measuring the same kinds of error. However, the largest correlation was r = 0.89, with most of the other correlations falling between r = .50 and r = .80. As a result, it is more probable that p-value correlation and RMSE are tapping into at least two different types of measurement error. As stated above, RMSE is rarely reported for the Angoff standard setting, and instead a large body of understanding about the p-value correlation has developed. A wider use of RMSE could lead to a better understanding of the Angoff standard setting and how judges assign scores to items.

In addition, the p-value correlations and the RMSE describe a slightly different situation for the listening data. Table 4.5 contains the only correlations from the two tests of convergent validity that do not reach significance. The correlation of RMSE 1 and p-value correlation is only r = -.22, the final correlation for RMSE 1 and RMSE 3 is r = .37, and the final correlation for RMSE 3 and p-value correlation 2 is r = - .37. These correlations hold particular importance because they illustrate the usefulness of the RMSE. Without the RMSE, the matrix would have consisted of only correlations of p-value correlations, and hence would have been composed of only significant correlations. The inclusion of RMSE as a measure of error and its performance in Table 4.5 illustrates that not only is RMSE a useful measure, it is useful precisely because it is measuring a kind of error that is not illustrated when using the p-value correlation.

It is also possible that some of the difference between Table 4.5 and Table 4.4 can be traced to differences in procedures used for the reading and for the listening panels. As described earlier, the listening panel did not actually get to listen to the questions as students would get to listen to the items when they are used on an actual test form. Judges read the questions off a script and then imagined how difficult the questions would be if students had to listen to them. Even though the judges were very familiar with how the test was constructed and presented to students, there could remain some confusion about how to assign difficulty estimates to items presented in this unconventional manner.

Table 4.5

Matrix Correlation for Measures of Listening Ability

	p-corr 1	p-corr 2	p-corr 3	RMSE 1	RMSE 2	RMSE 3	CSJ 1	CSJ 2	CSJ
p-corr 1	1.00	.44*	.60*	-.56*	-.38	-.48*	0.04	0.34	0.34
p-corr 2	X	1.00	.71*	-.22	-.48*	-.37	0.17	0.36	0.35
p-corr 3	X	X	1.00	-.56*	-.26	-.44*	0.01	0.26	0.32
RMSE 1	X	X	X	1.00	-.56*	.37	0.06	-0.13	-0.17
RMSE 2	X	X	X	X	1.00	.65*	0.15	0.11	0.20
RMSE 3	X	X	X	X	X	1.00	0.55*	0.37	0.19
CSJ 1	X	X	X	X	X	X	1.00	0.84*	0.54*
CSJ 2	X	X	X	X	X	X	X	1.00	0.84*
CSJ 3	X	X	X	X	X	X	X	X	1.00

**indicates p<.05*

Within-Judges Correlations

This last series of tables raises the issue that there might be different kinds of correlations that reveal different kinds of results about the data. A standard setting is a repeated measures design. The same subjects (in our case the judges) were measured several different times (three) on the same variables (judge's estimate of Θj). In the example of the standard setting in this study, the accuracy of the judge's estimates of Θj are measured by their correlation with the item p-value under high stakes conditions. In such designs, correlations can be decomposed into between-judges correlations and within-judges correlations.

The conventional measure in a standard setting of the p-value correlation can be thought of as a between-judges estimate of the judge's rating accuracy.

Between-persons (or between-subjects) effects…examine differences between individuals. This can be between groups of cases when the independent variable (IV) is categorical or between individuals when the (IV) is continuous. These type of effects can be observed in either the univariate context or the multivariate context (including repeated measures). Either way, between-subjects effects determine if respondents differ on the dependent variable (DV), depending on their group (males vs. females, young vs. old…etc) or depending on their score on a particular continuous IV. (Taylor, 2014).

As such, the correlations reported above in Tables 4.3, 4.4, and 4.5 are between-judges correlations. While important, between-judges effects alone ignore valuable information about the individual judges that are contained in the correlations obtained in the standard setting. Within-judges correlations are,

…[*within-subject] effects that represent the variability of a particular value for individuals in a sample. You see this commonly examined in repeated measures analysis (such as repeated measures ANOVA, repeated measures ANCOVA, repeated measures MANOVA or MANCOVA…etc). In these instances, a within person effect is a measure of how much an individual in your sample tends to change (or vary) over time. In other words, it is the mean of the change for the average individual case in your sample.* (Taylor, 2014).

As mentioned above, between-judges correlations of a judge's performance in an Angoff standard setting are the regular kind of p-value correlations gathered during a standard setting (Cizek, 1996; Cizek, 2001; Cizek & Bunch, 2007; Cizek, Bunch & Koons, 2004; Cizek, 2012). Within-judges correlations of performance in an Angoff standard setting have to be conceptualized differently. Rather than correlating estimates with the p-value of the item under high stakes conditions, judge's estimates should be correlated with their own estimates from different time periods in the same standard setting. So for example, a Pearson correlation of estimates between Round 1 and Round 2 will produce a comparison of whether or not the judge's estimates changed from Round 1 to Round 2, in a way that allows for comparison with other judges on the same or other items. A Pearson correlation of estimates between Round 2 and Round 3 will produce the same type of within-judge comparison of judges between those two rounds.

In this standard setting, the expectation is that judges will use the feedback information to help them formulate an accurate estimate of item difficulty, which in turn helps them position their estimate of the borderline student and the cutscore. This is done primarily after judges have made their Round 1 estimates but before they have made their Round 2 estimates. Following from this, judges should be changing their estimates between Rounds 1 and 2, and within-judges

correlations for judges at this time should be low. Within-judges correlations for Round 2 and 3 should be higher, as judges have already had access to feedback information and have developed a more stable image of the cutscore.

Tables 4.6 and 4.7 contain the reading and listening within-judges p-value correlations and squared residual scores for each of the judges in this standard setting. Within-judges correlations stand for whether the judges changed their correlation. If the correlation coefficient is high, this indicates that judges changed few of their estimates and presumably have a stable image of Θ_j. If the correlation coefficient is low, the judges have changed many of their estimates, and may be having trouble estimating b_i, thus producing an unstable estimate of Θ_j. In Table 4.6 for the reading panel, the column of results labeled, "Correlation between R1 and R2 using p-value correlation", is the Pearson correlation of the Round 1 Judge X's p-value correlation reading estimates with the Round 2 reading estimates of the same Judge X, and so on. The column of results labeled "Correlation between R1 and R2 using squared residual" is the Pearson correlation of the Round 1 Judge X's squared residual of the predicted value minus the squared residual of the observed value of Round 2 Judges X's reading estimates, and so on.

Tables 4.6 and 4.7 show the within-judges correlations for the judge's estimates between Rounds 1 and 2 and Rounds 2 and 3. For an N = 40 ($r = .30$, $p < .05$).

For the reading panel within-judges correlations (Table 4.6), the correlations between R1 and R2 are all significant except for two correlations from Panel 2. However, all correlations between R2 and R3 including those for Panel 2, are significant. In a comparison of the Group Means, at r = 0.57, Panel 2 had the lowest Group Mean during the Round 1 and 2 comparison. This was not true for the squared residual comparison. In the Round 1 and Round 2 comparison,

the Group Mean for Panel 3, at .60, was lower than for Panel 2, at .71. However, even though the Group Mean for Panel 2 increased substantially to .80, by the Round 2 to Round 3 comparison, with Panel 1 at .91 and Panel 3 at .85, Panel 2 had become the lowest Group Mean of all three panels.

For the listening panel (Table 4.7) a slightly different pattern occurred. All correlations in the table are significant. However, almost all of the judges showed lower Round 1 to Round 2 within-judges correlations, indicating that there was some problem with their initial understanding of item difficulty and hence their estimate of Θ_j. But by the Round 2 to Round 3 comparison, within-judges correlation had stabilized and the high Round 2 to Round 3 correlations indicate that less change in estimates was occurring.

Panel 2 Judges Agf 24 and Agf 25 were the only exception to this. These judges' estimates seem to indicate confusion about item difficulty that continued throughout the entire listening and the reading panels. In the reading panel, Panel 2 began with a within-judges correlation of p-value correlations for Round 1 and Round 2 that was lower than the other two panels, although the range of scores for these was very small. Similarly, for the correlation between the squared residual scores of Round 1 and Round 2, Panel 2 had the lowest Group Mean and remained the lowest throughout the standard setting. For the reading panel, it appears that Panel 2 correlations between Round 2 and Round 3 have reached a similar level of performance to that reached by Panel 1 and Panel 3 during their estimates from Round 1 to Round 2.

Similar results appear for Table 4.7 showing within-judges correlations for the listening panels. Panel 2, and particularly judges Agf 24 and Agf 25, began the standard setting with

extremely poor performance. By the end of standard setting, their performance had reached a level comparable to other panels during the Round 1 to Round 2 estimates.

Table 4.6

Within-judges Correlation between Estimates for p-value

Correlations and the Squared Residual Value - Reading

	Correlation between R1 and R2 using p-value correlation	Correlation between R2 and R3 using p-value correlation	Correlation between R1 and R2 using squared residual	Correlation between R2 and R3 using squared residual
Panel 1				
Agf11	.71*	.96*	.89*	.96*
Agf12	.73*	.95*	.74*	.85*
Agf13	.92*	.88*	.91*	.85*
Agf14	.77*	1.00*	.91*	1.00*
Agf15	.55*	.94*	.64*	.86*
Agf16	.69*	.82*	.83*	.94*
Group Mean	.73	.93	.82	.91
Panel 2				
Agf21	.67*	.92*	.81*	.90*
Agf22	.86*	.87*	.97*	.97*
Agf23	.84*	.94*	.88*	.95*
Agf24	.14	.45*	.41*	.61*
Agf25	-.03	.66*	-.31*	.46*
Agf26	.92*	.87*	.93*	.90*
Group Mean	.57	.79	.71	.80
Panel 3				
Agf31	.98*	.87*	.98*	.95*
Agf32	.95*	.81*	.96*	.78*
Agf33	.70*	.87*	.38*	.60*
Agf34	.93*	.96*	.88*	.93*
Agf35	.58*	.94*	.88*	.94*
Agf36	.74*	.96*	.40*	.92*
Group Mean	.81	.90	.75	.85

**indicates p<.05*

Table 4.7

Within-judges Correlation between Estimates for p-value

Correlations and the Squared Residual Value - Listening

	Correlation between R1 and R2 using p-value correlation	Correlation between R2 and R3 using p-value correlation	Correlation between R1 and R2 using squared residual	Correlation between R2 and R3 using squared residual
Panel 1				
Agf11	.79*	.96*	.88*	.95*
Agf12	.86*	.94*	.92*	.71*
Agf13	.65*	.96*	.87*	.93*
Agf14	.78*	.95*	.82*	.92*
Agf15	.70*	1.00*	.89*	1.00*
Agf16	.91*	.69*	.96*	.88*
Group Mean	.78	.92	.89	0.90
Panel 2				
Agf21	.62*	.86*	.65*	.84*
Agf22	.95*	.90*	.98*	.96*
Agf23	.76*	.78*	.70*	.87*
Agf24	.86*	.80*	.63*	.68*
Agf25	.33*	.68*	.38*	.45*
Agf26	1.00*	.95*	1.00*	.93*
Group Mean	.75	.83	.72	.79
Panel 3				
Agf31	.97*	.98*	.99*	.99*
Agf32	.90*	1.00*	.94*	1.00*
Agf33	.82*	.94*	.82*	.88*
Agf34	.93*	.98*	.94*	.97*
Agf35	.54*	.99*	.78*	.98*
Agf36	0.58*	1.00*	0.86*	1.00*
Group Mean	.79	.98	.89	.97

**indicates p<.05*

It appears that most judges were not able to initially understand item difficulty properly, and as a result, following the introduction of feedback data and discussion, they changed estimates for a large number of items. This is indicated by the low within-judge correlations reported for both the p-value correlation scores and the squared residual scores. However, with the introduction of feedback information after Round 1, the idea of item difficulty had become clearer and the within-judges correlations became higher indicating less changing of scores occurring between the period of Round 2 to Round 3. In addition, there appears to be group differences between the three panels on their ability to respond to feedback information. Panels 1 and 3 responded as expected for both the reading and listening panels and by Round 3, their estimates had become stable. Panel 2 was unable to respond to feedback in the same manner, and judges in this panel were still changing many scores even into Round 3, indicating a continued confusion about item difficulty and the borderline cutscore.

2. Do self-report measures of familiarity and confidence with one's knowledge of procedures and materials work effectively to predict judge's ability?

It is common practice to ask standard setting judges about their familiarity with and confidence in using the procedures of the standard setting they are following and also with their final results, as well as key aspects along the way to the final cutscore decision (Cizek, 2012; Hambleton et al., 2012; Loomis, 2012). During the standard setting used in this study, a series of self-report surveys were administered to monitor how the judges themselves perceived their progress, and how much confidence they had in their cutscore decisions.

This standard setting used a series of three different self-report surveys during Day 1 training and another two self-report surveys on Day 2 with one before and one after the operational standard setting (See Appendixes 6, 7, 8, 9 and 10). Table 4.8 shows the results of a correlation for the 18 judges on the 5 self-report surveys. Each result is shown for the two different measures of judge accuracy (i.e. p-value correlation and RMSE). The results of this correlation are only shown for Round 3. Rounds 1 and 2 were not shown because the judge's estimates were not finalized yet and were still being refined based on discussion and input information from the moderators. As we know from results reported above, in Rounds 1 and 2, judges are still exploring the items and often their estimates of scores can change considerably from Round 1 and Round 2, even between Round 2 and Round 3.

Table 4.8 shows the results of the correlation between judge's accuracy and their responses to the five different self-report surveys of familiarity and confidence. Day 2 measures reach significance for both reading and listening. The CSJ measures, while noticeably larger than those for the other measures, were still smaller than the Day 2 correlations and only one of the three

measurements on Day 1 was significant. In addition, CSJ was the only measure of accuracy to show significance for both the morning self-report and the final self-report measure.

These findings suggest that none of the training done on Day 1 had a large effect on the operational standard setting that occurred on Day 2. In fact, the correlations between Day 1 activities and Day 2 accuracy are so low, they seem to indicate that Day 1 training activities could be included effectively in a one-day training program, rather than the two-day event conducted for this study.

Table 4.8

Correlation of Measures of Judge Accuracy and Self-report Surveys for Round 3

	Assessment 4 Knowledge of Procedures	Assessment 5 Knowledge of CEFR	Assessment 6 Knowledge of PE Testing	Morning Day 2 Overview of Knowledge	Final Evaluation Confidence with Stan Set.
Reading					
p-value corr		.05	-.01	.32	.47*
RMSE		.20	.18	-.18	-.29
CSJ	.24	.40	.19	.42	.38
Listening					
p-value corr		.00	.14	.15	.62*
RMSE		.37	.11	.06	-.09
CSJ	0.42	0.45*	0.34	0.59*	0.66*

*indicates p<.05

CHAPTER 5 CONCLUSIONS & DISCUSSION

This study used data gathered from items in an Angoff standard setting to link an in-house examination developed by The University with the Common European Framework of Reference. The data consisted of the p-values of the original items used in the examination, as well as judge's estimates of what the p-values would be. Judge's p-value correlations and RMSE were also calculated where appropriate.

The research was conducted with two questions in mind, can judge's ability at estimating the true p-value of an item be predicted from (1) knowledge and training in Performance Level Descriptors (PLDs), and (2) self-report measures of familiarity and confidence with one's knowledge of procedures and materials? The results of this inquiry are reviewed below.

5.1 Summary of Results

Did knowledge of the PLDs predict a judge's final ability? No evidence was found in this study for such a relationship. Some of the judges who showed strong knowledge of the PLDs were among the lowest in demonstrated ability. Likewise, some of the judges who did not score well on the PLD Test were among the best performing judges in terms of their final ability. The idea that there is a relationship between knowledge of the PLDs and ability to determine cutscores is rejected.

This observation is especially sensitive. This study used the CEFR as its Performance Level Descriptors. The CEFR is widely used by teachers and researchers throughout the World. One of the findings of this study was the poor psychometric performance of the CEFR-based PLD Test. The PLD Test was composed of descriptors taken directly from the CEFR. It was only with difficulty that these descriptors could be used into this study. While the Council of Europe warns

against using these descriptors as a measure of language acquisition, describing them instead as statements of language competence, it is not entirely clear why the two ideas should be as different as found in this study. Because of this problem with the CEFR descriptors, they should only be used cautiously in research, and measures of reliability and suitability as a scale, such as Cronbach's Alpha, should also be reported. This problem suggested future research using the CEFR descriptors as measurement items should develop a proper measurement instrument rather than the list of items used in this study.

Do self-report measures of familiarity and confidence with the standard setting procedures predict a judge's accuracy? This study found a mixed response to this question. Measures taken on Day 1 of the study showed only a marginal relationship to judge's final ability. On Day 2, however, measures of familiarity and confidence showed a much better relationship with judge's final ability.

These findings can be interpreted that self-report measures are only of limited value when predicting judge's final performance in an Angoff standard setting. However, another interpretation is possible. The two-day standard setting that was described in this paper is unusual in that respect. Most standard settings are only one-day events. It may be no accident that the results of the first day of the two-day standard setting all show only marginal relationship with final performance. This is true not only for the self-report measures but also for the PLD Test, which was also held on Day 1. It may be that the poor results correlating the PLD Test with judge's final performance resulted not from the poor performance of the PLD Test, but rather because the standard setting was held over two days and the PLD Test was administered on Day 1.

Establishing whether the best length of time for a standard setting should be one or two days long would involve experimental manipulation of the length of training. Such a study is beyond the resources of this project, however, the optimal length of training remains a question of interest.

Another factor that may have played a hidden role in masking the effects of PLD knowledge and self-report results are the two related factors of social influence and decision-making styles. These factors were mentioned several times in the text as reasonable explanations for some of the differences that appeared in the study. The potential role for social influences is large, especially during the periods of training and standard setting before judges have developed a strong image of the borderline candidate.

Research on social influences in the standard setting has been generally ignored in more recent research work on the subject which has focused on the role of feedback and the way in which panelists use this to make their own estimates. It is clear from the results reported in this study that factors other than knowledge of the procedures of standard setting are responsible for the final performance of a standard setting judge. Social influences may play an important role, at least in some aspects of standard setting. In this study, judge failure was clustered inside panels and seemed to stem from problems with groups of individuals within a panel; rather than lone judges or panels. This is not necessarily the case with every standard setting panel, or even every panel that fails, but it certainly raises the issue. A revival of research into the social influences that play a role in standard setting may help overcome some of the problems produced by the small number of the panelists involved.

Despite these problems with the standard setting, the Angoff method worked as planned. Most judges changed their judgments during the interval between Round 1 and Round 2 after they had been exposed to the feedback data which included information about item difficulty. This gave judges the information they needed to position their estimates about the borderline candidates. Judges who were unable to do this appeared to have problems understanding the instructions and performing their duties as judges. Interviews with judges after they had performed their standard setting indicated that they had used the feedback information as intended – to help determine the difference between their first estimates and true estimates of the ability of the borderline student.

5.2 Other Important Findings

In addition to the results of the research questions, a number of other findings were established during this research.

The Listening Panels

The listening panel results did not exactly mirror the results of the reading panel. This was particularly true for the CSJ measure which was the only judge accuracy measure failing to reach significance in the repeated measures ANOVA. This may be because the listening panel and the reading panel were conducted quite differently. The reading panel was conducted in the conventional manner suggested by authoritative sources. Judges were presented with reading material and items exactly as the students would see them on a test.

The listening panel, on the other hand, was not conducted under similar conditions. During the training of the judges, listening items were broadcast to the judges, just as they would during a regular test. However, the actual operational standard setting itself was held in a different facility that did not have a self-contained broadcasting system. As a result, the listening

items were presented to the judges on a script. They were then asked to imagine what it is like to hear these items and then asked to use this imagination to estimate the difficulty of the items. This is quite different from the recommended test conditions for conducting a listening panel. A number of anomalous findings from the listening panel compared with the reading panel appear attributable to this problem.

The issue of the listening panel points to a larger problem with Angoff standard settings. Judges need to be presented with items under conditions as close as possible to the actual situation in which students would see them. Any attempt to present items in a manner different from this runs the risk of introducing irregularities making it difficult for the judges to produce answers that can be interpreted in the fashion necessary for a standard setting.

CSJ and RMSE in the Angoff Standard Setting

As mentioned above, the Root Mean Square of the Error or RMSE and the Cut-off Score Judgment are rarely used in interpreting standard setting results. Despite its usefulness here, there is not as much understanding about its role in the standard setting as there is for the more conventional p-value correlation. A clarification of these measures in standard setting could be useful to a better picture of what goes on in a standard setting and hence, the suitability of the cutscore decision.

Examples of where this is an issue are Table 4.4 and 4.5 of the correlation matrix between the p-value correlations and the RMSE for each of the rounds of the standard setting. In these tables, the p-value correlation indicates a straightforward relationship of the different measures of judge performance across the three rounds. The RMSE and CSJ, on the other hand, do not indicate a significant relationship with some of the p-value correlations. This seems to

indicate that RMSE is tapping into some kind of error that is different from the sort of error associated with the p-value correlation. This is not surprising given that p-value correlation is a measure of the relative differences between the predicted and observed variables, while the RMSE is a measure of the absolute difference in error. The p-value correlation is much easier to calculate and easier for standard setting organizers to interpret than is the RMSE or CSJ. This probably explains its widespread use; however it does not answer the question of what exactly these other measures are measuring in terms of the error of the standard setting. Further study of this could help clarify the issue.

5.3 Future Research Directions

Despite the failure of standard setting to reach full status as a psychometric technique, its use here raises several interesting points that could yield fruitful research results.

One possible research opportunity suggested by this research is that of locating activities that can be used successfully in the training to prepare standard setting judges. Much of what is used today to prepare standard setting judges appears to be based on its face validity and not in the measured efficacy of the activity. This is completely unacceptable for a procedure that holds such an important role in the test design process. Locating activities that really do provide panelists with the kind of skills and information they need to perform a successful standard setting is essential to improving the procedure. Very little is known about what these training procedures could be and the necessity to uncover such activities is both important and extensive.

One of the most difficult aspects of the method for the judges to handle appears to be using the PLDs to enhance their image of the ability and performance of the borderline student. Aside from the PLD Test, no attempt was made in this study to help judges use this information

other than by asking them to match descriptors. One of the more useful directions for further research could be an exploration of this problem. Clarifying the problem about the source of the poor correlation between judge performance and PLD Test score would be a necessary first step. Is this a problem of the poor psychometric properties of the current PLD Test? Or is it a problem of the extra time that this study used to train judges (two-days vs one-day)? Or is it simply that better activities need to be devised? A better understanding of issues such as these could be helpful in improving the performance of standard setting judges for the Angoff method or even other methods of determining cut scores.

5.4 Limitations of the Present Study

The present study examined whether or not standard procedures for training and preparing judges for the Angoff standard setting method worked to accurately predict judge's final performance. As with all research, the outcome of this study was limited in its findings by a range of factors.

Standard setting is an inherently weak form of research. Although research is necessary to investigate the procedures and results of standard setting, standard setting is a policy procedure organized to inform policy makers, rather than researchers, about a test. Given the small number of individuals involved, conventional statistical analysis and measurement is limited. It is very difficult to extract information about the effects that are the aim of the study, and about other important subjects such as gender, ethnicity or personal experience. In this study, an attempt was made to control for this by making panels as homogeneous as possible, while still maintaining representativeness. Despite these efforts, in the standard setting, researchers can never be certain that research results are not influenced by, or even the result of outside factors.

One of the solutions to this issue was to set the significance level for this study at a low level, the $p < .05$ level. But this carried with it its own set of risks. This paper reported on the results of 392 different correlation calculations. In addition, there were many exploratory correlations conducted but not reported. Setting significance at $p < .05$ was necessary because it allowed for the larger patterns of the standard setting to be observed. It also made it possible for many of the observations to occur by chance. At least 19 of the correlations reported in this study are expected to be significant by chance. Which ones these are, is of course unknown, but, while the trend in the results of this study are clear, individual findings need to be interpreted with caution.

The small size of the panels involved in standard setting have made it difficult to apply classical test theory, even though panels of this size - or smaller - are not uncommon in high stakes testing, such as those found in No Child Left Behind (Michigan State Department of Education, 2007). Attempts at adopting item response theory (IRT) to standard setting have gone largely ignored and as a result, standard setting has remained outside the domain of psychometrics continuing to involve an almost artistic sense to interpreting the tests.

Finally, while all the participants were selected because of their extensive experience in language teaching, this does not guarantee that the standard setting went as planned. The results of a standard setting are limited by the same training and skill that qualify the participants to be part of the standard setting. It is possible that some of the findings of this standard setting resulted from the biases and prejudges previously held by the judges. The recommended style of training judges for the standard setting is designed with this in mind and focuses directly on familiarizing judges with the CEFR and its proper use. Judges were asked several times about their understanding of the CEFR. However, this is no guarantee that training worked as intended.

It is possible that judges entered the standard setting using a standard different from the CEFR, and that the findings reflect a standard other than the one intended by the organizers and moderators. However, risk of this was kept to a minimum. On the morning of Day 2, just before the operational standard setting, judges were asked if they understood the CEFR. On a scale of 1 to 4, the mean average for all 18 judges was 3.06. So while this is a possibility, the judges indicated that they did not believe this to be a problem.

Despite careful planning, results also reflect the skills and abilities of the organizers and moderators of the standard setting event. Standard setting is a difficult and complicated process. The decisions it is used to manage are often very important. In the case of the standard setting from this study, it is possible that findings also reflect the limitations in the experience of the organizers and moderators.

REFERENCES

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (2014). Standards for educational and psychological testing. Amer Educational Research Assn.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In: R. L. Thorndike (Ed.), *Educational Measurement* (pp. 508-600). Washington, DC: American Council on `Education.

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education, 17*(1), 59–88.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch Model: Fundamental Measurement in Human Sciences.* Mahwah, NJ: Erlbaum.

Buckendahl, C. W., Smith, R. W., Impara, J. C., & Plake, B. S. (2002). A comparison of Angoff and Bookmark standard setting methods. *Journal of Educational Measurement*, *39*(3), 253-263.

Chi, M. T. H., Glaser, R., & Farr, M. J. (Eds.). (1988). *The Nature of Expertise*. Hillsdale, NJ: Erlbaum.

Cizek, G. J. (1996). An NCME instructional module on setting passing scores. *Educational Measurement: Issues and Practice*, *15*(2), 20-31.

Cizek, G. J. (2001). Conjectures on the rise and call of standard setting: An introduction to context and practice. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*, (pp. 3-17). Routledge.

Cizek, G. J. (Ed.). (2012). *Setting performance standards: Foundations, methods, and*

innovations. Mahwah. NJ: Erlbaum.

Cizek, G. J. (2012a). The forms and functions of evaluations in the standard setting process. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations,* (pp. 165-178). NJ: Erlbaum.

Cizek, G. J., & Bunch, M. B. (2007). *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests*. Thousand Oaks, CA: Sage.

Cizek, G.J., Bunch, M.B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice*, *23*(4), 31–50.

Clauser, J. C., Margolis, M. J., & Clauser, B. E. (2014). An examination of the replicability of Angoff standard setting results within a generalizability theory framework. *Journal of Educational Measurement*, *51*(2), 127-140.

Clauser, B. E., Mee, J., Baldwin, S. G., Margolis, M. J., & Dillon, G. F. (2009). Judges' use of examinee performance data in an Angoff standard-setting exercise for a medical licensing examination: An experimental study. *Journal of Educational Measurement*, *46*(4), 390-407.

Clauser, B. E., Mee, J., & Margolis, M. J. (2013). The effect of data format on integration of performance data into Angoff judgments. *International Journal of Testing*, *13*(1), 65-85.

Clauser, B. E., Swanson, D. B., & Harik, P. (2002).A multivariate generalizability analysis of the impact of training and examinee performance information on judgments made in an Angoff-style standard-setting procedure. *Journal of Educational Measurement*, *39*(4), 269–290.

Council of Europe. (2001). *Common European framework of reference for languages*. Cambridge: Cambridge University Press.

Council of Europe. (2009). *Manual for relating language examinations to the Common Common European Framework of References for Language Learning, Teaching, Assessment.* Cambridge: Cambridge: Cambridge University Press.

Crocker, L. & Zieky, M. (1994). *Joint Conference Standard Setting for Large-Scale Assessments*. National Assessment Governing Board. Washington, D.C.

Cronbach, L. J. (1988). Five perspectives on validation argument. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

Cross, L. H., Impara, J. C., Frary, R. B., & Jarger, R. M. (1984). A comparison of three methods on the National Teacher Examination. *Journal of Educational Measurement*, *21*(2), 113-129.

Egan, S. J., Dick, M., & Allen, P. J. (2012). An experimental investigation of standard setting in clinical perfectionism. *Behaviour Change*, *29*(3), 183-195.

Elman, B. A. (2000). *A cultural history of civil examinations in late imperial China.* University of California Press.

Embretson, S.E. and Reise, S.P. (2000). *Item response theory for psychologists.* Mahwah, NY: Lawrence Erlbaum Associates.

Engelhard, G. (2007). Evaluating bookmark judgments. *Rasch measurement Transactions*, 21, 1097-1098.

Engelhard, G. and Anderson, D. W. (1998). A binomial trials model for examining the ratings of standard setting judges. *Applied Measurement in Education*, *11*(3), 209-230.

Fitzpatrick, A. R. (1989). Social influences in standard setting: The effects of social interaction on group judgments. *Review of Educational Research, 59*(3), 315-328.

George, S., Haque, M. S., & Oyebode, F. (2006). Standard setting: comparison of two methods. *BMC Medical Education, 6*(1), 46.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist, 18*(8), 519–522.

Goodwin, L.D. (1999). Relations between observed item difficulty levels and Angoff minimum passing levels for a group of minimally competent examinees. *Applied Measurement in Education, 12*(1), 13-28.

Green, D. R., Trimble, C. S., & Lewis, D. M. (2003). Interpreting the results of three different standard setting procedures. *Educational Measurement: Issues and Practice, 22*(1), 22–32.

Halpin, G., Sigmon, G., & Halpin, G. (1983). Minimum competency standards set by three divergent groups of raters using a three judgmental procedures: *Educational and Psychological Measurement, 47*(1), 977-983.

Hambleton, R. K. (1980). Test score validity and standard-setting methods. *Criterion-referenced measurement: The state of the art*, 80, 123.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In Cizek G. J. (Ed.), *Setting performance standards: Concepts, methods, and perspectives*, (pp. 89-116).

Hambleton, R. K., Pitoniak, M. J., & Copella, J. M. (2012). Essential steps in setting performance standards on educational tests and strategies for assessing the reliability of results. In Cizek G. J. (Ed.), *Setting performance standards: Foundations, methods, and innovations (2nd ed., pp. 47–76). New York, NY: Routledge.*

Pitoniak, M. J. (2006). Setting performance standards. *Educational Measurement*, 4, 433-470.

Hertz, N. R., & Chinn, R. N. (2002, April). *The role of deliberation style in standard setting for licensing and certification examinations*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Holden, R. (2010). "Face validity". In Weiner, Irving B.; Craighead, W. Edward. (Eds,),

The Corsini Encyclopedia of Psychology (4th ed).(pp. 637-638). Hoboken, New Jersey: Wiley.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679-688.

Hurtz, G. M., & Auerbach, M. A. (2003). A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. *Educational and Psychological Measurement*, *63*(4), 584–601

Huynh, H. & Schneider, C. (2005). Vertically moderated standards: Background, assumptions, and practices. *Applied Measurement in Education*, *18*(1), 99-113.

Impara, J.C., & Plake, B.S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard-setting method. *Journal of Educational Measurement*, *35*(1), 69-81.

Jaeger, R. M. (1991). Selection of judges for standard-setting. *Educational Measurement: Issues and Practice*, *10*(2), 3-14.

Johnson, E. J. (1988). Expertise and decision under uncertainty: Performance and process. In M. Chi, R. Glaser, & M. J. Farr (Eds.), *The Nature of Expertise*. (pp. 209-228). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kaftandjieva, F. (2010). Methods for Setting Cut Scores in Criterion-references Achievement Tests. A Comparative Analysis of Six Recent Methods with an Application to Tests of

Reading in EFL. EALTA publication. Retrieved March 25, 2013 from http://www.ealta.eu.org/documents/resources/FK_second_doctorate.pdf

Kane, M. T. (2006). Validation. *Educational Measurement*, *4*(2), 17-64.

Kane, M. T. (2001). So much remains the same: conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: concepts, methods and perspectives* (pp. 19–51). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527-535.

Larkin, J. H., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Expert and novice performance in solving physics problems. *Science, 208,* 1335-1342.

Lavallee, J. (2012). Validation Issues in an Angoff Standard Setting: A Facets-based investigation. Unpublished PhD Dissertation, Department of Counseling and Educational Psychology, National Taiwan Normal University, Taipei, Taiwan.

Linn, R. L. (2003). Accountability: Responsibility and reasonable expectations. *Educational Researcher*, 32, 3-13.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher, 31,* 3–16.

Linn, R. L., & Shepard, L. A. (1997). Item-by-item standard setting: Misinterpretations of judge's intentions due to less than perfect item inter-correlations. In *Council of Chief State School Officers National Conference on Large Scale Assessment*, Colorado Springs, CO.

Lissitz, R. W. & Huynh H. (2003). Vertical equating for state assessments: Issues and solutions in determination of adequate yearly progress and school accountability. *Practical*

Assessment, Research & Evaluation, 8(10). Retrieved March 25, 2012 From http://pareonline.net/getvn.asp?v=8&n=10

Lissitz, R. W. & Wei, H. (2008).Consistency of standard setting in an augmented state testing system. *Educational Measurement*, *27*(2), 46-56.

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

Loomis, S. C. (2012). Selecting and Training Standard Setting Participants. *Setting performance standards: Foundations, methods, and innovations*, 107-134.

Lorge, L, & Kruglov, L. (1953). A suggested technique for the improvement of difficulty prediction of test items. *Educational and Psychological Measurement, 12*(4), 554-561.

McGinty, D. (2005). Illuminating the "Black Box" of standard setting: An exploratory qualitative study. *Applied Measurement in Education*, *18*(3), 269–287.

Margolis, M. J., & Clauser, B. E. (2014). The Impact of Examinee Performance Information on Judges' Cut Scores in Modified Angoff Standard-Setting Exercises. *Educational Measurement: Issues and Practice*, *33*(1), 15-22.

Mee, J., Clauser, B. E., & Margolis, M. J. (2013). The impact of process instructions on judges' use of examinee performance data in Angoff standard setting exercises. *Educational Measurement: Issues and Practice*, *32*(3), 27-35.

Messick, S. (1981). Constructs and their vicissitudes in educational and psychological measurement. *Psychological Bulletin*, *89*(3), 575–588.

Messick, S. (1989).Validity. In R. L. Linn (Ed.), *Educational measurement* (pp.13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research, 45*(1-3), 35–44.

Michigan State Department of Education. (February, 2007). Retrieved from http://www.michigan.gov/documents/mde/MI-ELPA_Tech_Report_final_199596_7.pdf

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The Bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249-281). Mahwah, NJ: Erlbaum.

National Council for Measurement in Education. (2015). Retrieved from http://www.ncme.org/ncme/NCME/Resource_Center/Glossary/NCME/Resource_Center/Glossary1.aspx?hkey=4bb87415-44dc-4088-9ed9-e8515326a061#anchorV

Nedelvsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14*(2), 3-19.

Nelson, D. S., (1994). Job analysis for licensure and certification exams: science or politics? *Educational Measurement: Issues and Practice, 13*(3), 29-35.

Norcini, J., Lipner, R., Langdon, L., & Strecker, C. (1987). A comparison of three variations on a standard-setting method. *Journal of Educational Measurement, 24*(1), 56-64.

Norcini, J. J. & Shea, J. A. (1997). The credibility and comparability of standards. *Applied Measurement in Education, 10*(1), 39–59.

Plake, B., & Giraud, G. (1998). *Effect of a modified Angoff strategy for obtaining item performance estimates in a standard setting study.* Paper presented at the Annual Meeting of the American Educational Research Association. San Diego, Calf.

Plake, B. S., Melican, G. J., & Mills, C. N. (1991). Factors Influencing Intrajudge Consistency During Standard-Setting. *Educational Measurement: Issues and Practice, 10*(2), 15-16.

Raymond, M. R., & Reid, J. B. (2001). Who made thee a judge? Selecting and training Participants for standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives*, (pp. 119-157).

Reckase M. D.(2000). The ACT/NAGB standard setting process: How "modified" does it have to be before it is no longer a modified-Angoff process? Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Reckase, M. D. (2006) Rejoinder: Evaluating standard setting methods using error models proposed by Schulz. *Educational Measurement*, *25*(3), 14-17.

Roach, A. T., McGrath, D., Wixon, C., & Talapatra, D. (2010). Aligning an early childhood assessment to state kindergarten content standards: application of a nationally recognized alignment framework. *Educational Measurement: Issues and Practice*, *29*(1), 25-37.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*(2), 413-428.

Schafer, W. D. (2005). Criteria for standard setting from the sponsor's perspective. *Applied Measurement in Education*, *18*(1), 61-81.

Schoonheim-Klein, M., Muijtjens, A., Habets, L., Manogue, M., Van Der Vleuten, C., & Van der Velden, U. (2009). Who will pass the dental OSCE? Comparison of the Angoff and the borderline regression standard setting methods. *European Journal of Dental Education*, *13*(3), 162-171.

Shepard, L.A. (1980). Standard setting issues and methods. *Applied Psychological Measurement*, *4*(4), 447-467.

Shepard, L. A. (1994). Implications for standard setting of the National Academy of Educational Evaluation of the National Assessment of Educational Progress achievement levels. In:

Proceedings of the joint conference on standard setting for large-scale assessments of the National Assessment Governing Board and the National Center for Educational Statistics (pp. 143–159). Washington, DC: U.S. Government Printing Office.

Smith, R. L. and Smith, J. S. (1988). Differential use of item information by judges using Angoff and Nedelsky procedures. *Journal of Educational Measurement*, *25*(4), 259-274.

Taube, K.T. (1997). The incorporation of empirical item difficulty data in the Angoff standard-setting procedure. *Evaluation and the Health Professions*, *20*(4), 479-498.

Taylor, J. (2014, July 17). Difference Between Within-Subject and Between-Subject [Blog] Retrieved from http://www.statsmakemecry.com/smmctheblog/within-subject-and-between-subject-effects-wanting-ice-cream.html

van de Watering, G., & van der Rijt, J. (2006). Teachers' and students' perceptions of assessments: A review and a study into the ability and accuracy of estimating the difficulty levels of assessment items. *Educational Research Review*, *1*(2), 133-147.

Verhoeven, B. H., Verwijnen, G. M., Muijtjens, A. M. M., Scherpbier, A. J. J. A., & Van der Vleuten, C. P. M. (2002). Panel expertise for an Angoff standard setting procedure in progress testing: item writers compared to recently graduated students. *Medical Education*, *36*(9), 860-867.

Wessen, C. (2010). Analysis of Pre- and Post-Discussion Angoff ratings for evidence of social influence effects. Unpublished MA Dissertation, Department of Psychology, University of California, Sacramento.

Wiley, A., & Guille, R. (2002). *The occasion effect for "at-home" Angoff ratings*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Yin, P. & Schultz, E. M. (2005). *A comparison of cut scores and cut score variability from Angoff-based and Bookmark-based procedures in standard setting.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Appendix 1. Common European Framework of Reference - Global Scale

Level	Performance Level Descriptors
C2	Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.
C1	Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.
B2	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without

	strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and Independent disadvantages of various options.
B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.
A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.

A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

Source: CoE, 2001, p. 24.

Appendix 2. Informed Consent Form

Informed Consent Form for ELC Standard Setting Pilot Studies (July 2010)

The ELC assessment committee is doing research on the standard setting process used to link tests to the Common European Framework of Reference (CEFR). The results of this process are important, because they determine what test scores count as 'proof' that a student has reached a certain ability level. However, the process itself is very subjective, and there is no way to prove that a given score means that a student has "really" reached a given ability level. *The purpose of this study is thus to help us better understand the factors that influence the decision-making process, as part of the longer-term goal of improving the process.*

You are being invited to take part because of your background in the TEFL field. Your participation is entirely voluntary, and your choice will have no bearing on your job or on any work-related evaluations or reports. If you accept, you will be asked to complete a short preparatory assignment and to participate in two one-day workshops. At these meetings, you will receive more training and then you will be asked to make a series of judgments concerning the difficulty level of items from The University's English exams in relation to the CEFR descriptors.

Audio and video recordings will be made of the group discussion and the interviews, and the recordings will be transcribed. The recordings will be treated as confidential and no real names will be used in the transcripts. No one outside of the Assessment Committee will have access to the transcripts.

We expect that your participation will help us to better understand how judges make standard setting decisions. We will share any findings with you. The findings may also be shared with other researchers in the field through presentations and/or publications.

All participants will receive an honorarium.

If you have any questions about any aspect of the study, please do ask.

Joseph Lavallee, Principal Investigator

ELC Assessment Committee, The University

I have read the foregoing information. I have had the opportunity to ask questions about it and any questions I have been asked have been answered to my satisfaction. I consent voluntarily to participate in this study.

Name (please print clearly): _____

Signature: _____

Date (day/month/year): _____

Appendix 3. Security Form

Standard Setting Security Agreement Form

I, _____ (print panelist name here), understand and accept the following terms and conditions.

1. Panelists will follow all test security procedures set forth in writing or verbally by English Language Center representatives.

2. Panelists will turn over to English Language Center representatives all products of the standard setting meeting at the close of the session or as directed by said representatives.

3. Panelists relinquish any claim or right to any and all products turned over to the English Language Center.

4. All of the materials used at the meeting are considered secure and panelists are expected to turn in all such materials at the close of the session or as directed by English Language Center representatives and to maintain complete confidentiality regarding these materials.

_____ _____

Panelist signature Date

Appendix 4. Angoff Panelist Record Form

PANELIST NAME: _____ Listening - ROUND 1

Circle or **insert** the probability that a just-B1 level student would get the item correct. Then write your probability at the bottom of the table.

	ITEM NO.							
	1	2	3	4	5	6	7	8
	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4
	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6
	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7

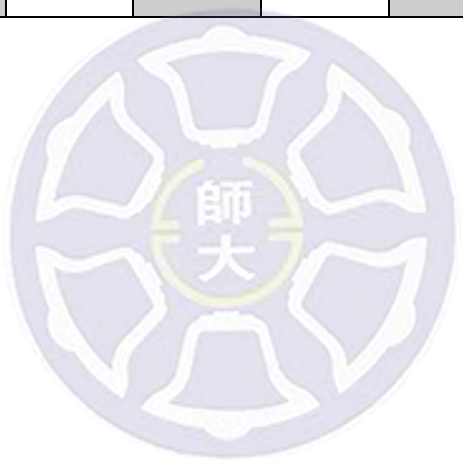
	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9
Prob.								

PANELIST NAME: _____ Reading - ROUND 1

Circle or **insert** the probability that a just-B1 level student would get the item correct. Then write your probability at the bottom of the table.

	ITEM NO.							
	1	2	3	4	5	6	7	8
	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4
	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6

	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.7
	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9
Prob.								



Appendix 5. Panelist Information Form

Panelist Information Form

Please Answer the Following Questions

Name (in native language)

_____

Current Job Title/Employer

_____

Highest relevant degree attained (please list the name of the degree and the granting institution):
_____

_____

Years of teaching experience with university-level students in Taiwan:

_____

Years of experience with test design/development:

_____

Have you ever lived in an English-speaking country? Where? How long?

_____

Please circle the answer that best shows your opinion

I am familiar with the test design and construction process.

Disagree 1 2 3 4 Agree

I am familiar with the Common European Framework of Reference (CEFR).

Disagree 1 2 3 4 Agree

I am familiar with Rasch or IRT Modeling.

Disagree 1 2 3 4 Agree

I am familiar with the standard setting process.

Disagree 1 2 3 4 Agree

Appendix 6.

PART I. Procedures Panelist Number _____

Please circle the answer that best shows your opinion

I have completed the introduction to the procedures of a standard setting.

Disagree 1 2 3 4 Agree

The group leader answered all of my questions.

Disagree 1 2 3 4 Agree

I understand the instructions so far.

Disagree 1 2 3 4 Agree

I understand why a standard setting is important.

Disagree 1 2 3 4 Agree

I have made decisions about student standards before.

Disagree 1 2 3 4 Agree

I feel qualified to make this kind of judgment.

Disagree 1 2 3 4 Agree

I feel I have the experience to make this kind of judgment.

Disagree 1 2 3 4 Agree

Standard setting is common sense.

Disagree 1 2 3 4 Agree

Appendix 7.

PART II. Common European Framework

Panelist Number _____

I have completed the introduction to the CEFR.

Disagree 1 2 3 4 Agree

The group leader answered all of my questions.

Disagree 1 2 3 4 Agree

I understand the instructions so far.

Disagree 1 2 3 4 Agree

I understand the CEFR.

Disagree 1 2 3 4 Agree

I understand the difference between the different levels of the CEF.

Disagree 1 2 3 4 Agree

I understand the B1 level.

Disagree 1 2 3 4 Agree

The CEFR is a useful way to think about teaching English.

Disagree 1 2 3 4 Agree

I agree with the order of the CEFR levels.

Disagree 1 2 3 4 Agree

The CEFR is common sense.

Disagree 1 2 3 4 Agree

Appendix 8.

PART III. The University Practical English Test

Panelist Number _____

I have completed the introduction to the Practical English Test.

Disagree 1 2 3 4 Agree

The group leader answered all of my questions.

Disagree 1 2 3 4 Agree

The practice test helped me understand more about the test.

Disagree 1 2 3 4 Agree

I understand the instructions so far.

Disagree 1 2 3 4 Agree

I understand item difficulty.

Disagree 1 2 3 4 Agree

I understand item proficiency scales.

Disagree 1 2 3 4 Agree

I understand scale scores.

Disagree 1 2 3 4 Agree

I understand why some items are more difficult than others.

Disagree 1 2 3 4 Agree

Appendix 9.

Review of Standard Setting Procedures

I have completed Angoff Standard Setting training.

Disagree 1 2 3 4 Agree

I understand the Angoff Standard Setting

Disagree 1 2 3 4 Agree

I understand the procedures of an Angoff Standard Setting.

Disagree 1 2 3 4 Agree

I understand the Common European Framework.

Disagree 1 2 3 4 Agree

I understand item difficulty.

Disagree 1 2 3 4 Agree

I feel qualified to perform an Angoff Standard Setting.

Disagree 1 2 3 4 Agree

I understand the B1 level of the CEFR.

Disagree 1 2 3 4 Agree

I feel ready to start the Angoff Standard Setting.

Disagree 1 2 3 4 Agree

I am ready to begin the Angoff Standard Setting.

Appendix 10.

Angoff Standard Setting. Final Evaluation page 1 of 2

Panelist Number _____

The training and practice exercises helped me understand how to perform the task.

Disagree 1 2 3 4 Agree

The time provided for discussions was adequate.

Disagree 1 2 3 4 Agree

There was an equal opportunity for everyone in my group to contribute his/her ideas and opinions.

Disagree 1 2 3 4 Agree

I was able to follow the instructions and complete the evaluation accurately.

Disagree 1 2 3 4 Agree

The discussions after the first round of ratings were helpful to me.

Disagree 1 2 3 4 Agree

The discussions after the second round of ratings were helpful to me

Disagree 1 2 3 4 Agree

I am confident about the defensibility and appropriateness of the final recommended cutscores.

Disagree 1 2 3 4 Agree

The facilities and food service helped create a productive and efficient working environment.

Disagree 1 2 3 4 Agree

Angoff Standard Setting. Final Evaluation page 2 of 2

Panelist Number _____

The information showing the distribution of student scores was helpful to me.

Disagree 1 2 3 4 Agree

I found the discussion between rounds to be useful.

Disagree 1 2 3 4 Agree

I changed my scores between rounds.

Disagree 1 2 3 4 Agree

The discussion between rounds influenced me to change my score.

Disagree 1 2 3 4 Agree

The information about student performance influenced me to change my score.

Disagree 1 2 3 4 Agree

I am confident with my final decisions.

Disagree 1 2 3 4 Agree

I believe my final decisions are the best decision I can make.

Disagree 1 2 3 4 Agree

Appendix 11

Cutscore statistics for the Standard Setting - reading

Judge	Round 1	Round 2	Round 3
Agf11	85.28	65.38	75.25
Agf12	71.25	74.35	74.80
Agf13	77.00	76.08	77.88
Agf14	76.75	74.25	79.00
Agf15	69.63	79.00	75.38
Agf16	72.38	70.13	80.38
Agf21	53.90	81.63	56.70
Agf22	81.35	57.03	81.48
Agf23	71.00	79.35	76.25
Agf24	70.43	74.75	67.38
Agf25	44.50	70.38	68.88
Agf26	71.63	75.38	73.00
Agf31	75.63	72.38	75.38
Agf32	73.50	76.00	76.00
Agf33	63.50	74.50	75.13
Agf34	69.25	71.13	74.58
Agf135	62.83	72.95	76.68
Agf136	64.75	76.43	78.23
Mean	69.70	73.39	74.58
SD	9.52	5.56	5.65

Appendix 12

Cutscore statistics for the Standard Setting – listening

Judge	Round 1	Round 2	Round 3
Agf11	79.00	77.13	76.75
Agf12	78.63	75.25	73.88
Agf13	85.15	81.30	79.18
Agf14	77.13	77.75	80.38
Agf15	79.25	78.75	78.63
Agf16	76.50	78.25	82.50
Agf21	66.88	68.75	68.38
Agf22	81.50	80.88	80.88
Agf23	71.75	78.50	86.75
Agf24	53.63	57.88	68.75
Agf1525	74.18	63.38	66.25
Agf26	73.88	73.88	72.38
Agf31	76.13	76.63	76.88
Agf32	79.50	79.88	79.88
Agf33	74.88	77.88	77.38
Agf34	79.75	79.13	78.13
Agf135	76.63	80.35	81.30
Agf136	83.88	81.63	81.50
Mean	76.01	75.96	77.21
SD	7.04	6.42	5.40