國立台灣師範大學教育心理與輔導學系

博士論文

指導教授：林世華 博士

以多面向Rasch模式為基礎檢驗Angoff標準設定法的效度議題

(Validation Issues in an Angoff Standard Setting: A Facets-based Investigation)

研究生：李炯方 撰 (JOSEPH P. LAVALLEE)

中華民國一百一年六月

ACKNOWLEDGMENTS

I would like to express my gratitude to my advisor, 林世華 (Lin Sieh-Hwa), for his guidance at every stage of this long process. In my many visits to his office, he always shared not only technical suggestions about my thesis but also generous doses of wisdom and perspective. I am grateful also to the additional members of my committee, 林邦傑 (Lin Pang-Chieh), 鄭夙珍 (Nellie Cheng), 陳柏熹 (Chen Po-Hsi) and 張武昌 (Vincent Chang), whose thoughtful comments have made this a better dissertation than it otherwise would have been.

I thank my friends, Scott Sommers, 洪素蘋 (Hung Su-Ping) and 黃宏宇 (Huang Hung-Yu), for all of the long and stimulating discussions about educational measurement, and my friend Quentin Brand for occasionally reminding me that other topics merited discussion as well.

林瑩玲 (Grace Lin), 詹雨臻 (Chan Yu-Chen), 林小慧 (Lin Shiao-Hui) and 方威傑 (Johnny Fang) have all gone well out of their way in the past few years to help me to navigate my way around linguistic and other barriers as I've worked my way through the program; my 'sister' Deborah Kraklow kindly offered comments on various drafts of this dissertation to help make it more readable. I thank all of them for giving so generously of their time.

Finally, I thank Brian Lin for his patient support and quiet encouragement.

以多面向Rasch模式為基礎檢驗Angoff標準設定法的效度議題

李炯方（JOSEPH P. LAVALLEE）

摘要

近年來，標準設定方法在教育實務情境中蓬勃發展，其中尤以修正版Angoff標準設定法的使用最為廣泛。Angoff法假定，經過訓練後的評分者能依據試題難度正確地估計出通過預設標準的最低能力受試者，其答對每一道試題的成功機率。由於標準設定方法的主觀評分特性，因此，尋求適切的工具以確保評分者評分品質甚為重要。多面向Rasch模式(MFRM)已被廣泛使用於主觀評分情境，特別是在標準設定程序中，用以考驗評分過程中是否出現負向的評分者效果而影響評分品質。然而，多面向Rasch模式的基本假設為，評分者間的影響是不存在的。然而由於多數的研究除了評分資料外並未能取得相對客觀的試題難度資料加以比對以考驗此假設，因此極少有研究檢驗該假設。由於使用Angoff法時，除了評分者對於試題難度的評估以及受試者是否有能力能夠達到預先設定的標準，同時還可以取得外部試題反應資料。基於此，本研究利用Angoff法所取得的外部試題反應資料以及評分者資料，來交叉驗證多面向Rasch模式的基本假設。其次，利用多面向Rasch模式來檢驗Angoff法的三個假設，以及評分資料與模式的適切程度。

在執行Angoff法時，研究者請18位外語教學(EFL)專家擔任評分者，並將英文閱讀以及聽力試題各40題對照到歐洲語言共同架構中的B1等級(Common European Framework of Reference)。在負向評分者效果的偵測方面，本研究依據MFRM所提供的各項指標，偵測三種在評分過程常出現的評分者效果：嚴苛度(leniency/severity)、準確度(inaccuracy)以及趨中與極端評分 (centrality/extremism)。接著，將Angoff設定法所估計的概率作為內在參照架構，並將施測所得的試題難度估計作為外在參照架構。首先，將MFRM指標用來偵測在兩個參照架構下的評分者效果，並比較兩個架構下標準設定的結果。其次，利用原始分數以及MFRM指標來考驗Angoff標準設定法的基本假定。

本研究主要的發現如下：

1. 對照兩個架構下的標準設定，評分者在嚴苛度、準確度以及評分趨中與極端程度的結果不一致。如此的差異使研究者對於單獨使用Angoff設定法，作為設定標準分數的方式，產生疑慮。有關群體效果假設的考驗也確實發現，在使用內部的參造架構下，確實出現群體趨中評分效果。這也顯示出在使用多面向Rasch模式前必須先考驗評分者間的群體效果是否存在。
2. 關於Angoff法的假設檢定，BPS以及試題功能方面違反基本假設。其中較嚴重的缺失為，幾乎所有的評分者皆無法利用概率來評估最低受試者能力。

關鍵字：標準設定、Angoff法、多面向Rasch模式、評分者效果、歐洲語言共同架構、評分品質

Validation Issues in an Angoff Standard Setting: A Facets-based Investigation

JOSEPH P. LAVALLEE（李炯方）

Abstract

Introduction: The use of standards-based scores in education has grown in recent years and the modified Angoff standard setting method is perhaps the most widely used procedure for establishing these standards. In this method, trained judges imagine students who just meet the standard in question and estimate the likelihood of their responding correctly to each item on the test being aligned to the standard. The method assumes that trained judges can accurately represent students who just meet the standard, represent how test items function and quantify their estimation of the likelihood of student success for each item. All three assumptions have been called into question. More generally, the subjective nature of all standard setting methods has resulted in a focused search for tools to evaluate the quality of judges' decisions.

The many-facet Rasch model (MFRM) has been proposed for use in detecting rater effects generally and for evaluating standard setting results in particular. Use of the MFRM, however, relies on the further assumption that no group-level rater effects exist. Because only internal, judge-generated data is available in most cases, this assumption is usually not evaluated and little research exists on how plausible the assumption is in real settings or on how robust results are to violations of the assumption. As external item response information often is available when the Angoff method is used, an Angoff setting provides a rare opportunity to test this assumption of the MFRM. Thus, the two-fold purpose of this study is to first evaluate the suitability of the many-facet Rasch model using data from an Angoff standard setting, and then to evaluate the assumptions of the Angoff method using the MFRM.

Method: The data consisted of the first round estimates of a panel of 18 trained EFL professionals serving as judges in an operational Angoff standard setting linking two

40-item English exams (one reading, one listening) to the Common European Framework of Reference B1 proficiency level, and of the item response data from the original administration of the exams. MFRM indices were identified for the detection of three broad types of rater effects: leniency/severity, inaccuracy and centrality/ extremism. These indices include estimated parameters and standard errors, residuals and residual-based indices, separation statistics and correlations between ratings and model indices. The probability estimates made by the Angoff judges were used to construct an 'internal' frame of reference, and the item difficulty estimates from the test administration were used to construct an 'external' frame of reference. Indices from the many-facet Rasch model were used to examine the subjective ratings of the Angoff judges for the presence of rater effects in both frames and the results were compared. In the second stage of the study, the assumptions of the modified Angoff method were assessed, using raw score and MFRM indices.

Results: In the first phase, results differed across frames for all three rater effects. The leniency/severity indicators suggested greater agreement between judges in the internal frame than in the external frame, although a similar number of judges were flagged (four in both the internal and external frames for reading; two in the internal and three in the external frame for listening). Inaccuracy effects were sharply underestimated within the internal frame of reference: six judges were flagged in the internal frame and nine in the external frame for reading; for the listening test, two and four judges were flagged in the internal and extermal frames respectively. Results for centrality/extremity differed even more markedly: for the reading test, four judges were flagged for centrality and five for extremism in the internal frame while 17 judges were flagged for centrality in the external frame; for the listening test, 10 judges were flagged for centrality and one judge for extremity in the internal frame while all 18 judges were flagged for centrality in the external frame. Group-level indicators did indicate the presence of group-level centrality and inaccuracy effects within the internal frame of reference, suggesting their possible use in evaluating the assumption of the model prior to use.

In terms of the assumptions of the Angoff method, the BPS and item functioning assumptions appear to have been violated to some extent but the most striking failure was the inability of nearly all judges to accurately quantify their assessments using the probability scale. The 'centrality' or 'central tendency' bias, in

particular, was displayed by nearly all judges, compressing the Angoff metric. This compression of the scale appears to have been largely responsible for the distorted results for the MFRM leniency/severity and centrality/extremity indices in the internal frame noted above. Further, this scale compression appears to have distorted the cut scores, leading to differences in pass/fail rates: for the reading test, the pass rates within the internal frame across the three rounds of the standard setting were 46.4%, 37.8% and 37.7%, while the corresponding pass rates in the external frame were 38.1%, 29.0% and 27.2%; for the listening test, the pass rates in the internal frame were 35.4%, 35.4% and 31.5%, compared to 31.0%, 31.0% and 27.1% in the external frame.

Discussion: The critical assumption underlying use of the MFRM for detecting rater effects was found not to hold in the present case, casting doubt on the use of the model in standard setting situations for which only internal data (from the judges' estimates) is available. More positively, the group-level indicators within the internal frame were found to be sensitive to inaccuracy and centrality effects and thus may serve to help check the suitability of the model for use where no external data is available.

The assumptions of the Angoff method were also found to be violated. In particular, a centrality or central tendency bias was shown to persist across all three rounds and to distort results. In view of previous research into central tendency, the present findings are consistent with the possibility that the Angoff method is inherently highly susceptible to the distorting effects of this bias. More generally, the centrality bias seems likely to pose a serious threat in many rating situations, both to the validity of ratings *and to the accuracy of indicators used to evaluate these ratings*.

Future research should focus on refining our understanding of when the MFRM is likely to be appropriate for use; on solutions to problems with the Angoff method (perhaps in the form of procedural modifications or score adjustments); and on what rating situations are likely to be susceptible to the centrality bias and how it might be reduced or eliminated.

Keywords: standard setting, Angoff method, many-facet Rasch model, rater effects, Common European Framework of Reference, rating quality

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 INTRODUCTION

1.1 Significance of the Current Study

In recent years, the use of standards-based scores has become increasingly widespread, internationally as well as in Taiwan. When significant consequences are attached to meeting these standards, validity becomes an issue of obvious importance. Since most standard setting methods rely on subjective judgments made by content-area experts, assessing the validity of the results involves evaluating the quality of such judgments. However, methods for this task are themselves still under development. The two-fold purpose of the present study is to use data from a single operational standard setting to both empirically assess the methods used to evaluate subjectively-made judgments *and* to evaluate the assumptions of the standard setting method itself.

Standard setting methods are employed so that scores from an examination can be reported in relation to a standard. Several methodologies have been devised that make use of expert judgment to arrive at a numerical cut score linking an examination to a standard. The most commonly used standard setting methodology is the modified Angoff method (Angoff, 1971). In the modified Angoff standard setting method, a panel of content-area experts is trained to imagine a 'barely proficient student' who has *just* achieved the proficiency standard in question and then to work through the items in a test, estimating for each one the probability that the barely proficient student would answer it correctly. The estimates are summed and the average across raters is the recommended cut score. Typically, this is an iterative procedure involving two or three rounds, with empirical performance data from the test in question used to provide feedback when available.

The claim that the method generates valid results rests on a set of assumptions which minimally include that trained judges can: (1) develop accurate representations of the *just*-proficient examinee; (2) accurately represent item functioning (i.e., the features of each item that make it either more or less difficult for examinees); and (3) juxtapose these two representations to arrive at a quantitative estimate - the probability that the just-proficient examinee can respond correctly to the item. Each of

these three assumptions has been questioned in the literature and there are well-established reasons for believing that they may not hold in all, or even typical, cases.

Given these known difficulties, methods for assessing the quality of judges' decisions would clearly be of considerable value, and different diagnostic indicators have been suggested for use in evaluating the results of standard setting meetings. A growing body of literature has emerged around the use of the many-facet Rasch model (MFRM) and related latent trait models for detecting a number of rater effects, including leniency/severity, inaccuracy and central tendency. In recent years, a number of authors have proposed using this model for the detection of rater effects in the context of standard setting exercises (Eckes, 2009; Engelhard, 2007; 2009; 2011; Engelhard & Anderson, 1998; Englehard & Cramer, 1997; Engelhard & Gordon, 2000; Engelhard & Stone, 1998; Noor, 2007).

However, use of the MFRM in a standard setting context is itself based on a further assumption. Namely, it assumes that any rater effects are confined to a minority of raters and that no group-level effects are present. This is because the collective ratings of the group are used to define the model expected values, in relation to which deviations can be isolated and identified as particular rater effects. If group-level effects exist, they would influence the expected values themselves. Thus, the claim that the expected values can be used to detect rater effects depends on the assumption that, *at the level of the entire group of raters, there are no rater effects*.

Typically, it is very difficult to evaluate this assumption. In most rating situations, the only data available comes from the judgments of the raters themselves; there is no external data available against which it can be compared. Thus, despite the large and growing body of literature around the use of the MFRM for detecting rater effects, the viability of this assumption has rarely been questioned. From this perspective, an Angoff standard setting provides an unusual opportunity. In many situations in which the Angoff method is used, data on item difficulty exists from the original administration of the exams. These item difficulty parameters can be used to construct an external frame of reference, which can then be used to evaluate the assumption of the MFRM. The present study is the first to explicitly attempt to determine whether this assumption holds within a particular rating situation and, if the

assumption is not met, how robust results are when violations of the assumption occur.

The purpose of this study is thus two-fold. By constructing an external frame of reference from original test results and an internal frame of reference from expert judgments, it seeks to first evaluate the underlying assumption of the many-facet Rasch model, and then to use the MFRM, along with raw score indicators, to evaluate the assumptions of the Angoff standard setting method. The study thus seeks to contribute to our understanding of the use of the MRFM for detecting rater effects. In terms of standard setting procedures, it is hoped that the study will add to the literature exploring the characteristics and evaluating the assumptions of the Angoff method. Identification of specific threats to validity should be useful for refining the procedures used in future standard settings, the design of the training conducted prior to their implementation and for the validation process that occurs after the cut score has been established.

1.2 Research Questions

The analyses conducted in this study seek to address the following research questions:

1) Does the critical assumption required for the use of latent-trait models for the detection of rater effects in a criterion-referenced situation hold in a typical modified Angoff standard setting?

2) Are the assumptions of the modified Angoff standard setting viable?

a) Are trained Angoff panelists able to develop accurate representations of the 'barely proficient student'?

b) Are trained Angoff panelists able to develop accurate representations of item functioning?

c) Are trained Angoff panelists able to juxtapose these representations to assess the likelihood of the barely proficient student answering each item correctly and to quantify this assessment using the 0-1 probability scale?

1.3 Terminology

Angoff Method – Often referred to as the 'modified Angoff' method, this is a procedure for generating a cut score linking a particular exam to the achievement of a standard. In this procedure, a trained panel of subject-matter experts goes through the test and, for each item, estimates how many of a group of 100 barely proficient students (students who *just* meet the standard) would respond to it correctly. Results are summed for each panelists and averaged across the panel to arrive at the cut score. Often two or three rounds of judgments are conducted, with empirical data used to provided feedback to panelists between rounds, if such data is available.

Barely Proficient Student (BPS) – The (imaginary) student who just meets the standard in question. In the Angoff method, panelists are asked to develop an internal image of such a student after familiarization with the performance standard. The BPS is sometimes referred to as the '*minimally competent examinee*,' the '*borderline examinee*' or the '*just-proficient student*.' In this study, these terms, along with the *'B1 BPS'* and '*just-B1 student'* will be used interchangeably.

Common European Framework of Reference (CEFR) - A manual developed by the Council of Europe (CoE, 2001) to provide common reference materials for the teaching and learning of different languages. The manual contains 54 language proficiency scales, covering various aspects of language performance. The proficiency scales consist of six basic levels, labeled, in increasing order of proficiency, A1, A2, B1, B2, C1 and C2. The CEFR scales have been adopted for use internationally in providing 'performance standards.'

Cut score – The 'cut score' is the test score which translates between the performance standard and performance on the test. A student whose test score is at or above the cut score is said to have reached the performance standard.

Facilitator – A facilitator is a person responsible for conducting the training and the actual standard setting meeting.

Judges– All standard setting methods requiring subjective judgments require *panelists* or *judges* who will make these decisions under the guidance of the meeting facilitator(s). They are typically expected to be subject-matter experts and are also sometimes expected to be representative of different stakeholder groups.

Modified Angoff Method - See "Angoff Method", above.

Performance standards – Benchmarks against which performances can be measured.

Performance Level Descriptors (PLDs) – Descriptions of the characteristics of performance at a given level. In this study, the key 'PLDs' are the B1 level descriptors for listening and reading from the Common European Framework of Reference.

CHAPTER 2 LITERATURE REVIEW

Section 2.1 introduces the Angoff method and the assumptions that it makes, along with the different types of rater errors which could lead to violations of those assumptions. Section 2.2 introduces the *many-facet Rasch model (MFRM),* which has been proposed for use in detecting the presence of rater effects. The model and its indices are introduced and the use of the model for investigating rater effects is described. The central assumption made for this purpose is introduced and means for evaluating this assumptions are discussed.

2.1 The Angoff Method: Assumptions and Validity Threats

One of the most commonly procedures employed in standard setting is the modified Angoff method (Angoff, 1971; hereafter simply the "Angoff method"). With this method, a panel of judges, usually content-area experts, is trained to imagine a 'barely proficient student' (BPS) who has *just* achieved the proficiency standard in question. After being trained, the judges consider each item in the test, one by one, estimating for each the probability that the BPS would answer the item correctly. The sum of these estimates for each judge represents the 'cut score' recommended by that judge; that is, the score on the test that an examinee would need to reach to be considered as having reached the standard in question. The average cut score across the entire panel of judges is taken as the recommended cut score. Typically, this is an iterative procedure involving two or three rounds, with empirical performance data from the test in question used to provide feedback, where such data is available.

2.1.1 Assumptions of the Angoff Method

Use of a cut score to make potentially high-stakes decisions about examinees assumes that *if* students who perfectly exemplified the 'barely proficent' ability level for the standard took the test, *they would receive the same score as did the barely proficient students imagined by the Angoff judges.* This implies further assumptions. Angoff did not explicitly state these assumptions and no single 'list' is agreed upon in the various discussions of the procedure (Brandon, 2004; Impara, 1997; Impara & Plake, 1998; Ricker, 2006). Nonetheless, the method seems to assume, at a minimum, the

following:

1. Accurate Representation of the Barely Proficient Student. Trained judges can develop accurate representations of the ability level of the just-proficient student.

2. Accurate Representation of Item Functioning. Trained judges can accurately represent the nature and level of knowledge, skills and abilities required to respond to the item when making their estimates.

3. Quantification. Trained judges can juxtapose their representations of the BPS and of item functioning to assess the degree of challenge posed by the item for the BPS and quantify this using the 0-1 probability scale.

Threats to these assumptions, which would call the validity of Angoff results into question, are discussed next.

2.1.2 Known Threats to the Assumptions of the Angoff Method

The Angoff method is likely the most thoroughly researched of all standard setting methods. Findings of this research as they relate to the three assumptions listed above are summarized here.

Assumption 1: Accurate Representation of the Barely Proficient Student. In developing mental representations of the just-proficient student, panelists are internalizing the construct as it is articulated in the performance level descriptors (Bourque, 2000; Egan et al., 2009; Lewis & Green, 1997; Mercado & Egan, 2005). All verbal descriptions of ability are likely to leave some degree of ambiguity. The ambiguity of the CEFR descriptors used in the present study, for example, has been widely discussed (e.g., Weir, 2005). Given this, it may be more reasonable to think of a 'zone' within which accurate BPS representations might exist rather than a single point. Put differently, within limits, different experts or judges might have different but more or less equally defensible interpretations of the written standards. Thus, 'accurate BPS representations' are here understood as those which fall within a 'zone' along the latent trait continuum. Threats to validity exist when BPS representations fall *outside* of this range, such that they cannot be defended as reasonable interpretations or 'translations' of the PLDs.

Research on different types of performance level descriptors (PLDs) and the results have offered some support for the BPS assumption. Impara, Giraud & Plake

(2000) found judges set a higher cut score on the same exam when given PLDs reflecting a higher degree of proficiency. Skorupski and Hambleton (2005) found that teacher descriptions of performance levels converged after training and orientation activities. Giraud, Impara & Plake similarly found that teachers given more detailed PLDs generated more detailed descriptions of the BPS, and Fehrmann, Woehr, Arthur (1991) found that two groups of panelists who received more thorough training with practice rounds produced estimates that were in closer agreement than a third group which received minimal training.

The literature on standard setting has focused on three general variables which might lead to violations of the BPS assumption: factors other than the PLDs influencing the development of BPS representations, panelist background and panelist stakes in the outcome.

In a qualitative study, McGinty (2005) found that judges seemed to be basing their representations of the BPS on particular students who had been granted degrees (indicating achievement of the standard) instead of on the PLDs describing what students *should* master to earn the degree. Reid (1985) found having judges make estimates for the total group before doing so for the target group lowered the cut score for the target group considerably, suggesting the possibility that consideration of the student population as a whole influenced the development of the BPS representation. Similarly, in a study of a speaking and a listening test being linked to the CEFR, Papageorgiou (2010) found that some panelists relied on information other than the PLDs in making their judgments.

Another set of studies have focused on the performance of panelists with different backgrounds. Hamberlin (1992, in Brandon, 2004) found that non-teachers in a school (administrators, curriculum specialists, etc.) set significantly higher standards than did teachers. It may be that teachers, more familiar with the precarious nature of newly learned skills, developed a somewhat less 'able' BPS, or that administrators place a higher priority on setting higher standards that would reflect well on the school. Cross et al. (1984) found that public high school teachers and teacher-educators in universities set different cut scores on a teacher education battery. Busch and Jaeger (1990) found similar effects for public school and college/ university-based judges on a similar test, noting that ratings provided by the public

school judges were more influenced by item performance data than were the ratings from the college/university content specialists. Verhoeven et al. (1999, 2002) compared practicing professionals (doctors) with recently graduated students, and found that the recent graduates gave more homogeneous judgements and set a significantly more lenient cut score. Another study found that psychology graduate students setting cut scores for a psychology test had less variation in their scores than a group of undergraduate students who had just taken the course, suggesting that the graduate students may have shared more similar representations of the BPS (Maurer et al., 1991). However, other studies (e.g., Norcini, Shea, & Kanya, 1988; Plake, Impara & Potenza, 1994) have found no significant difference in the ratings provided by judges with different backgrounds.

McGinty (2005) also found that the consequences associated with different cut scores also seemed to contaminate the process, with judges who were high school teachers feeling a tension between the desire to set high standards and the desire to be viewed by the public as doing a good job (which would be called into question if they set a high cut score resulting in more students failing). In that study, the majority initially wanted to set high standards but, McGinty observed, "reality set in when some participants pointed out that teacher performance would be judged by the passing rates on the test." Consistent with McGinty's conclusion, Ferdous and Plake (2005) found that judges in the U.S. who indicated that they were influenced by the consequences of cut scores in relation to the No Child Left Behind law set lower cut scores.

In the present study, the PLDs were comparatively very detailed and the training period relatively lengthy. Further, it is unlikely that the panelists anticipated significant consequences resulting from their cut score decisions. These considerations would suggest a high level of agreement about the ability level of the BPS. On the other hand, the panelists came from rather diverse backgrounds, ranging from recently graduated students, to native English speaking teachers with years of classroom experience and panelists with administrative job positions. These background differences might be expected to produce divergent BPS representations.

Assumption 2: Accurate Representation of Item Functioning. This has surely been the most controversial and well-researched assumption of the Angoff method

(Brennan & Lockwood, 1980; Chang, 1999; Chang et al., 1996; Clauser et al., 2009; Fehrmann, Woehr & Arthur, 1991; Goodwin, 1999; Hurtz & Jones, 2009; Impara & Plake, 1998; Lorge & Kruglov, 1953; Plake & Impara, 2001; Plake, Impara & Irwin, 1999; Shepard, Glaser, Linn, & Bohrnstedt, 1993; Van Der Linden, 1982). The overwhelming consensus which has emerged from this research is that judges are indeed quite limited in their ability to represent item functioning. Most such studies have reported correlations between the means of modified Angoff judges' item estimates and actual difficulty levels (i.e., empirical *p*-values). Brandon's 2004 review of the literature on the Angoff method reported that, for 29 correlations reported, average correlations were .63 for operational standard settings and .51 for non-operational standard settings. This moderate level of success in meeting the assumption has remained the rule in studies published since Brandon's review (e.g., Clauser et al., 2009).

Research in this area has increasingly sought to investigate the variables influencing accuracy in assessing item difficulty. Panelist background and expertise has been the focus of one line of research, with inconclusive results. Van De Watering and Van Der Rijt (2006) found that students were more accurate than their teachers but the Verhoeven studies discussed above failed to find a difference between panelists with different backgrounds.

Assumption 3: Quantification. After developing representations of the BPS and of item functioning, Angoff judges next need to juxtapose these representations, imagine how the just-proficient student would interact with the item, conceptualize the degree of challenge posed by the task *and* 'quantify' this by estimating the probability of the BPS answering correctly. The ability of panelists to quantify their expectations as probabilities has rarely been explicitly discussed. This is curious, as there is little reason to expect this to be a natural task for most people and, conceptually, it is not clear how a panelist is expected to perform it.

Furthermore, previous research offers reason to believe that the *central tendency* or *centrality* effect, in particular, may commonly occur when the Angoff method is used. The centrality effect has long been known to influence judgments made in settings similar to that of the Angoff. Indeed, over a century ago, Hollingworth noted that judgments of "time, weight, force, brightness, extent of

movement, length, area, size of angles, have all shown the same tendency to gravitate toward a mean magnitude, the result being that stimuli above that point in the objective scale were underestimated and stimuli below overestimated" (Hollingworth, 1910, p. 426). This effect has been consistently found within the psychophysics tradition: Stevens and Greenbaum (1966) reviewed a series of experiments demonstrating the same effect, which they referred to as a "regression effect." More than a decade later, Poulton provided an updated review of the literature concerning this tendency, which he referred to as "contraction bias" and described as "a general characteristic of human behavior" (Poulton, 1979, p. 778). Unfortunately, this literature has rarely been referred to in relation to the Angoff method, despite its obvious relevance.

If the centrality effect were present in an Angoff setting, it would manifest as a tendency for judges to overestimate the difficulty of relatively easy items and to underestimate the difficulty of relatively difficult items. The standard deviation of judges' estimates would also be smaller than the standard deviation of the empirical item difficulties (i.e., those derived from the actual administration of the test to the relevant student population).

Precisely this pattern of results has been found in a number of studies. In Lorge and Kruglov's (1953) study of the ability of judges to estimate item difficulty, the standard deviation of the judges' estimates was 16.3 compared to 23.7 for the empirical item difficulties. Shepard (1994) found that trained Angoff judges systematically overestimated examinee performance on difficult items and underestimated examinee performance on easy items. In Goodwin (1999), 14 judges made estimates for all examinees and for the borderline examinees on a 140-item financial certification exam. The standard deviations of the judge's estimates were .09 for the total group and .10 for the borderline group; the corresponding standard deviations from the actual exam results were .19 and .18 respectively. Heldsinger and Humphry (2005) and Heldsinger (2006) reported results from a study in which 27 judges used a modified Angoff procedure with 35 items from a Year 7 reading test. The standard deviation of the item difficulties set by the panelists was 0.5 logits, less than half the standard deviation of 1.16 logits from the actual exam results. The authors used the ratio of the standard deviations to re-scale the Angoff results and

found that that it significantly altered the final cut score. Schulz (2006), in addition to providing one of the first attempts to theoretically elucidate the nature of this bias as it relates to standard setting, reported results from a pilot study, with 21 Angoff panelists making estimates for items from the 2005 NAEP Grade 12 math exam. The results suggested 'scale shrinkage' which, significantly, persisted even through the third round of ratings. Finally, Clauser et al. (2009), reported results from two operational standard setting exercises for a physician credentialing examination, with six Angoff judges making estimates for 200 items (34 of which has associated empirical data) on one, and six judges and 195 items (43 with empirical data) on the other. Even though items with "very high" or "very low" p-values were excluded from the study, the judges were still found to "systematically overestimate the probability of success on difficult items and underestimate the probability of success on easy items" (Clauser et al., 2009, p. 17).

In fact, results consistent with a centrality effect appear to have been found *every time they have been looked for*. The one seeming exception is a study by Impara and Plake, in which, according to the authors, panelists "did not systematically overestimate (or underestimate) performance on easy items or overestimate (or underestimate) performance on hard items" (Impara & Plake, 1998, p. 77). However, the particular methodology used in that study makes it difficult to directly compare their results with the studies mentioned above. In their study, the authors asked 26 sixth-grade science teachers to estimate the probabilities of success on each item in a 50-item science test for two groups: the borderline ("D/F") students in their class, and the class as a whole. They also asked the teachers to assign and record the class grades for each student. The researchers then compared predicted with actual performance for both groups *with the borderline group defined by the teacher-assigned class grades.* They found that the teachers overestimated the performance of the class as a whole but underestimated the performance of the borderline group. They then examined the relationship between predicted and actual item difficulty levels for both groups, categorizing estimates as *overestimates* (more than .10 over the actual p-value), underestimates (more than .10 under the p-value) and accurate estimates (within .10 of the actual p-value). The results were then further divided according to the difficulty level of the item (items with p-values below .34, between .34 and .66

and above .66). They concluded that "these results did not show a consistent variation in accuracy of prediction simply as a function of item difficulty" (p. 77).

This study certainly speaks to the ability of panelists to estimate the performance of particular students and may be of particular interest in comparing the modified Angoff method with student-centered standard setting methods, such as the contrasting groups method. Nonetheless, their results cannot be compared directly with results from the studies mentioned above, for at least two reasons. First, as noted in Clauser et al. (2009), Impara and Plake defined the borderline group *in terms of the class grades assigned by the teachers*. In order to make a direct comparison of estimated and observed difficulty levels, the authors would have needed to have defined the groups *statistically,* in accordance with the modified Angoff method: by the number of items each group was *predicted* to answered correctly. Doing so would have resulted in a different set of proportion-correct ('p') values, a different categorization of items into the three levels of item difficulty, and different percentages of estimates falling into each of the accuracy categories used by the authors (*overestimates, accurate estimates* and *underestimates*). In other words, the relevant comparison is with students who performed around the mean score derived from the teachers' item-by-item estimates. Second, the authors provide no information on the dispersion of estimates, such as the range or standard deviation. Without these, and given the above issue of category definition, Impara and Plake's findings cannot be used as evidence either for or against the presence of a centrality bias.

In short, then, based on previous research, there is strong reason to believe that the Angoff method is highly vulnerable to a central tendency bias which has the potential to undermine one of its core assumptions.

2.1.3 Rater Effects

An important part of the validation process generally is to identify possible threats to validity, formulate them as hypotheses and then seek to empirically refute them (APA/ AERA/NCME, 1999; Kane, 1994; Messick, 1998). For standard setting and subjective rating situations more broadly, such hypotheses can be explicitly formulated in terms of the presence of possible 'rater effects,' defined as a "broad category of effects [resulting in] systematic variance in performance ratings that is

associated in some way with the rater and not with the actual performance of the ratee" (Scullen, Mount and Goff, 2000, p. 957). These rater effects have been investigated in some depth within two broad research traditions. The first of these has focused on the psychological processes involved in making subjective evaluations and on the potential sources of rater effects (Pula & Huot, 1993). The second tradition has focused on detecting and diagnosing rater effects by searching for their characteristic patterns in ratings data. Research within this latter tradition has resulted in a variety of criteria to evaluate the psychometric quality of ratings, across different measurement frameworks, including classical test theory, analysis of variance, regression analysis, generalizability theory and Rasch measurement/item response theory (Saal, Downey, & Lahey, 1980; Stemler, 2004; Stemler & Tsai, 2008). Within this broad literature, rater effects have been defined in various ways (Myford & Wolfe, 2003, 2004; Saal, Downey & Lahey, 1980). The present study will follow Wolfe's division of rater effects into three categories leniency/severity, inaccuracy and centrality/extremism (Wolfe, 2004). These are discussed in turn.

Leniency/Severity. This effect is present when raters gives scores that are consistently either too high or too low. In terms of an Angoff standard setting, leniency/severity is present when a judge's probability estimates are uniformly either lower or higher than is warranted by the performance-level descriptors. A judge displaying a leniency bias would assign comparatively low probability estimates to the items, resulting in a lower cut score and a higher percentage of students meeting the standard. Conversely, a judge displaying a severity bias would attribute to the BPS more ability than warranted by the PLDs and would thus assign higher probability estimates, resulting in a higher cut score and a lower percentage of students meeting the standard.

Inaccuracy. To the extent that this effect is present, ratings will appear unrelated to the presence or absence of the latent trait being rated. In an Angoff standard setting, this effect would create inaccuracies in the representations of item functioning.

(It should be noted that, within the broader category of *inaccuracy*, it is possible to make a further distinction between *randomness* and *differential dimensionality* (Wolfe & McVay, 2011). Randomness is present when a rater's ratings

diverge in a *non-systematic* manner from error-free measurements (Wolfe, 2004), whereas *differential dimensionality* occurs when ratings *systematically* deviate from the ratings that would be assigned by an error-free process, violating the assumption of local independence and the related assumption of unidimensionality. Differential dimensionality may result due to a number of specific biases which have been discussed in the literature, such as *halo effect* (Saal et al., 1980, p. 474), *logical error* (Newcomb, 1931; Linn & Gronlund, 2000), or bias/interaction effects (Lumley & McNamara, 1995; Lynch and McNamara, 1998; Wigglesworth, 1993). However, pursuing and further specifying the cause of inaccurate ratings is beyond the scope of the present study.)

Centrality/Extremism. The centrality effect (discussed above in relation to the *quantification* assumption) is present when a rater clusters his or her ratings around a certain point of the rating scale or around the center of the perceived range of performances, resulting in a compressed distribution. This results in reduced variation in assigned ratings, and in ratings that are accurate at the center of the ability range but which overestimate the ability of less proficient examinees and underestimate the ability of more proficient examinees. In an Angoff setting, this would mean overestimation of the probability of success for more difficult items and underestimation of the probability of success for easier items.[1]

A less frequently discussed effect is *extremism,* present when ratings cluster at extreme ends of the rater's distribution of ratings (Wolfe, 2004). Where this effect is present in an Angoff setting, difficult items would tend be judged as being even more difficult than they really are, and vice-versa, easier items would be judged as being even easier than they actual are.

2.2 Detection of Rater Effects with the MFRM

In recent years, latent trait models, and the many-facet Rasch model (MFRM) in particular, have been widely proposed for use in detecting, diagnosing and, to some

---

[1]C*entrality* or *central tendency* is often defined to occur when ratings cluster near the midpoint of the rating scale, and distinguished from *range restriction* which is defined to occur when ratings cluster around any point of the rating scale (Saal, Downey & Lahey, 1980). Here, in line with an earlier tradition (Hollingsworth, 1910), these are treated as a single rater effect occurring when ratings cluster around the average rating.

extent, adjusting for rater effects (Eckes, 2005; Engelhard, 1992, 1994, 1996; Linacre, 1989; Myford & Wolfe, 2003, 2004). Of particular interest here, a number of researchers have applied the MFRM to detect rater effects in the context of standard setting (Eckes, 2009; Engelhard, 2007; 2009; 2011; Engelhard & Anderson, 1998; Englehard & Cramer, 1997; Engelhard & Gordon, 2000; Engelhard & Stone, 1998; Noor, 2007). This section first describes the relevant members of the Rasch family of models and discusses the indices which have been proposed for the detection of specific rater effects. This application of the model relies on the assumption that no group-level rater effects are present. This assumption is often left implicit and, to date, there is no instance in the literature in which it has been explicitly evaluated prior to application of the model. This assumption is thus treated at some length, and means for evaluating this assumption are described.

2.2.1 Latent Trait (Rasch) Models: Parameters and Indices

It may be possible to infer the presence of rater effects from the parameter estimates generated by the model, from the residuals between expected and observed values, from the separation statistics, and from the correlations between ratings and the indices generated by the model. Each is discussed below.

Latent Trait Models and Model Parameters

In Rasch's original model for dichotomous data, responses are a stochastic function of person and item parameters:

$$\ln(P_{ni1} / P_{ni0}) = \beta_n - \delta_i \qquad (2.1)$$

where β_n is the location of person n along the underlying latent trait, δ_i is the location of item i along the same latent variable, and P_{ni1} and P_{ni0} are the probabilities of person n on item i scoring 1 and 0, respectively. Applying the model assumes the existence of a quantitative underlying variable (e.g., EFL reading or listening ability), and when parameters are estimated from the raw response data, an interval scale along which all examinees and items can be located is generated for this variable. The

distance between any two items, any two students or any item and any student indicates a specific quantity of the attribute being measured. The origin of the scale is arbitrary (often set at the mean item difficulty location), as is the unit which partititions the latent variable into specific quantities. Such a situation describes a single *specified frame of reference*, understood as a collection of agents (students), a collection of objects (items), and outcomes of the interactions between them (Rasch, 1977).

The frame of reference for the dichotomous model is constructed by setting up interactions between agents and objects so as to transmit variations in underlying person ability and item difficulty to the measurement outcome, performance on the test. The Rasch *rating scale model* (Andrich, 1978) expands the dichotomous model to allow for polytomous cases, so that

$$\ln(P_{nijk} / P_{nij(k-1)}) = \beta_n - \delta_i - \tau_k \qquad (2.2)$$

where $\tau_{k,}$ refers to the threshold between two adjacent rating scale categories, and where P_{nijk} and $P_{nij(k-1)}$ refer to the probability that person *n* attempting item *i* endorses category *k* and category *k-1*, respectively.

Note that in many empirical frames of reference involving a rating scale, the relationship between the latent variable and measurement outcome has been dramatically changed. Variation is no longer transmitted through the direct interaction between person and item. Rather, ratings involve *subjective perception and judgment* concerning the outcome of the interaction between the task or item and the person. This opens the way for *rater effects* to influence the measurement outcome. The *many-facet Rasch model* (MFRM; Linacre, 1989) provides a further extension of the model to take account of other features or 'facets' of the rating situation, starting with the severity of the different raters. Thus, a typical facets model for a three-facet situation can be formulated as follows:

$$\ln(P_{nijk} / P_{nij(k-1)}) = \beta_n - \delta_i - \lambda_j - \tau_k \qquad (2.3)$$

where β_n is the ability of person n, δ_i is the difficulty of item i, λ_j is the severity of judge j and τ_k is the difficulty of observing category k relative to category k-1, and P_{nijk} and $P_{nij(k-1)}$ refer to the probabilities of examinee n being graded on item i by judge j, with a rating of category k and k-1, respectively.

In the above model, the rating scale is the same for all judges and all criteria. However, it is also possible to use a *partial-credit model* (Masters, 1982) in which each criterion has its own rating scale. This may be desirable if it is expected that the use of the scale categories is expected to differ between different criteria. This model can be formulated as follows:

$$\ln(P_{nijk} / P_{nij(k-1)}) = \beta_n - \delta_i - \lambda_j - \tau_{ik} \tag{2.4}$$

In this model, τ_k refers to the difficulty of observing category k on item i. In other situations, we may want the scale to remain the same across criteria but allow it to vary across judges. This can also be modeled using the following formula:

$$\ln(P_{nijk} / P_{nij(k-1)}) = \beta_n - \delta_i - \lambda_j - \tau_{jk} \tag{2.5}$$

In this model, the threshold of the steps between adjacent rating categories varies among judges. The two-subscript term, τ_{jk}, refers to the difficulty of observing category k used by judge j. Further facets can be included in the model based on hypothesized sources of construct-irrelevant variance. Thus, if interaction effects (e.g., between judges and criteria or raters and ratees) or sources of differential rater functioning (DRF) such as gender or professional background are believed to be present, these can also be included.

$$\ln(P_{nijk} / P_{nij(k-1)}) = \beta_n - \delta_i - \lambda_j - \gamma_g - \tau_k - \varphi_{jg} \tag{2.6}$$

where γ_g is the ratee group (e.g., gender, profession) and φ_{jg} is a bias interaction term representing the interaction between judges and a group of ratees. This facet indicates

the degree to which rater *j*'s ratings for ratee group *g* differ from the expected ratings of rater *j* for ratee group *g*, as predicted by a model *not* containing the term.

Another model which can be used for parameter estimation is the binomial trials model (Wright & Masters, 1982), which is used when the response format calls for a specific number of independent attempts at each item with a dichotomous outcome (success or failure), and the number of successes is counted. The responses are thus defined as the number of independent successes. This model is as follows:

$$\ln(P_{nix} / P_{nix-1}) = \beta_n - \delta_i - \tau_x \qquad (2.7)$$

where $\tau_{x,}$ refers to the difficulty of achieving a count of *x* relative to a count of x-1, and where P_{nix} and P_{nix-1} refer to the probability of person *n* achieving a count of *x* on item *i* and a count of *x-1* on item *i*, respectively. This model has been recommended as being particularly appropriate for a modified Angoff situation (Eckes, 2009; Engelhard & Anderson, 1998; J.M. Linacre, personal communication, July 7, 2010), since in the modified Angoff method, judges are presented with a series of dichotomous items and essentially asked: "Out of 100 barely proficient students, how many would answer this item correctly." The probabilities obtained can thus be modeled as outcomes of binomial trials, with the number of independent trials fixed at 100 and the judges asked to count the number of successes (i.e., the number of barely proficient students who would answer the item correctly).

Separation Statistics

The family of Rasch models generates a series of indices designed to ensure that the elements of a particular facet (e.g., judges or items in an Angoff setting) are "sufficiently well separated in difficulty to identify the direction and meaning of the variable" (Wright & Masters, 1982, p. 91). These indices depend on the standard errors of the parameter estimates and the standard deviation of the elements of the facet being analyzed.

The ratee separation ratio, *G*, is a measure of the spread of the ratee performance measures relative to their precision, with separation expressed as a ratio

of the 'true' (adjusted for measurement error) standard deviation of the measures over the average standard error. The standard error associated with each particular estimate is calculated as the square root of the maximum score ('*mN*') divided by the observed score ('*S*') multiplied by the maximum minus the observed score; the result is multiplied by a factor ('*Y*') which increases with sample dispersion to control for the spread of the sample.

$$SE = Y\sqrt{\frac{mN}{S(mN - S)}} \tag{2.8}$$

This formula inflates the denominator and results in lower standard errors for elements near the center of the distribution, which are more 'well-targeted'; the relatively less well-targeted elements at the extremes have larger standard errors.

The mean square error (MSE) is the mean of the error variances:

$$MSE = \frac{\sum_{i=1}^{N} SE_i^2}{N} \tag{2.9}$$

The MSE can be used to adjust the observed variance of the measurements. Because each individual measurement contains error, a 'true' variance can be defined as:

$$\text{True } SD^2 = SD^2 - MSE \tag{2.10}$$

The *average* error or *root mean square error* (RMSE) is then defined as the square root of the mean square error:

$$RMSE = \sqrt{MSE} \tag{2.11}$$

With these indices in place, an element separation ratio, *G*, can be computed using the adjusted or 'true' standard deviation and the root mean square error, as follows:

$$G = \text{True SD} / \text{RMSE} \tag{2.12}$$

This ratio is a measure of the spread of the measures relative to their precision. It can also be used to derive the separation index, *H*, which indicates the number of measurably different levels (or strata) of performance. This index is defined as:

$$H = (4G + 1) / 3 \tag{2.13}$$

The proportion of observed variance which is not due to estimation error is used to indicate the reliability with which the elements in the sample are separated:

$$R = \frac{SD_{True}}{SD_{Observed}} = 1 - \frac{MSE}{SD_{Observed}} = \frac{G^2}{1+G^2} \tag{2.14}$$

Finally, a chi-square statistic is generated by the Facets (MFRM) program to assess the statistical significance of the differences between the elements within each facet. For the rater or judge facet in an Angoff setting, the chi-square statistic is calculated as follows:

$$\chi^2 = \sum \left[\frac{\frac{\beta_n^2}{\sigma_n^2} - \left(\sum \frac{\beta_n^2}{\sigma_n^2} \right)^2}{\sum \frac{1}{\sigma_n^2}} \right] \tag{2.15}$$

Where β_n is the cut score for Judge N and σ_n is the associated standard error. The statistic has an approximate chi-square distribution with *N - 1* degrees of freedom, where *N* is the number of judges. The chi-square statistic for the items in an Angoff

can be calculated by substituting the corresponding values for the items into the equation.

Residuals-based Indices

The family of Rasch models uses the residuals between the expected ratings generated by the model's parameter estimates, and the actual observed ratings, to generate fit statistics which may help to indicate possibly mismeasured examinees, raters or tasks. The residual for an observation is defined as:

$$R_{nij} = X_{nij} - E_{nij} \tag{2.16}$$

where

X_{nij} = the observed rating, and

E_{nij} = the expected rating, based on model parameter estimates.

The formula for the expected rating, E_{nij}, is

$$E_{nij} = \sum_{k=0}^{m} kP_{nijk} \tag{2.17}$$

where

m = the number of rating scale categories, and

k = a counting index representing the value of each rating scale category.

Residuals are usually standardized for interpetation, using

$$Z_{R_{nij}} = \frac{R_{nij}}{\sqrt{V_{E_{nij}}}} \tag{2.18}$$

where

$$V_{E_{nij}} = \sum_{k=0}^{m} (k - E_{nij})^2 P_{nijk} \tag{2.19}$$

Here, note that $V_{E_{nij}}$ is the variance of an observation so that its square root is its statistical information.

Two fit statistics are generated for each parameter estimate, based on the mean of the squared standardized residuals of the observed scores from their expected scores. As they are based on model residuals, fit statistics capture and summarize deviations from expected ratings. The *outfit* statistic is simply the mean of these standardized residuals. Outfit statistics are particularly sensitive to departures in the data in the extreme rating categories. *Infit* statistics attach a weight to each standardized residual based on its variance, making them more sensitive to unexpected ratings that fall near the center of the rating scale. Infit and outfit statistics have an expected value of 1.00 and can range from zero to infinity. A 0.1 increase in a fit statistic is associated with a 10% increase in unmodeled error. Values less than 1.0 indicate that the model predicts the data better than expected, based on model expectations for levels of error.

Outfit is calculated as follows:

$$OutFit_{\lambda} = \frac{\sum_{n=1}^{N}\sum_{i=1}^{I} z_{rni}^{2}}{NI} \tag{2.20}$$

where, N is the number of examinees, I is the number of items and z_{rni}^{2} is the standardized score residual.

The formula for infit is:

$$InFit_{\lambda} = \frac{\sum_{n=1}^{N}\sum_{i=1}^{I} z_{rni}^{2} W_{rni}}{NI} \tag{2.21}$$

where W_{rni} is the variance of the score residual. Both of these statistics can be standardized to obtain the standardized infit and outfit statistics. According to a widely used rule of thumb for interpreting fit statistics, overfit to the model (better

than expected model-data fit) is suggested when mean square fit statistics fall below 0.7 for multiple-choice exams and 0.6 for rating situations, and misfit or underfit to the model (worse than expected model-data fit) is suggested when the values are above 1.3 for multiple-choice exams and 1.4 for rating situations (Wright & Linacre, 1994). The corresponding values for standardized fit statistics are ±2.0.

Correlational Indices

Many of the indices which have been recommended for use in detecting rater effects are correlations between different model-generated indices or between model indices and raw score ratings.

There are two widely used raw score correlations. The *'single rater-rest of rater'* (SR/ROR) correlation, which is the Facets version of the point-biserial correlation, is a raw score indicator derived by calculating the correlations between the ratings of the different raters across all facets. A second raw score indicator, not specific to the Facets model but widely used in Angoff settings, is the correlation between the estimates of the judges and the empirical item p-values. In this study, these are referred to as *score/p-value correlations*.

A number of latent trait correlations are also widely used as indicators. The *point-measure correlation* is the latent trait analog to the point-biserial correlation. It is the correlation between the scores assigned by a particular rater to a group of examinees and the ability estimates for the same examinees, and thus indicates the consistency between how a particular rater ranks the examinees and how the raters collectively rank the same group. Similarly, the *score-expected correlation* is the correlation between the observed and modeled scores. Finally, the *expected-residual correlation* is a measure of the relationship between the residual (observed minus modeled or expected score) and the modeled or expected scores themselves, while the closely related *measure-residual correlation* indicates the relationship between the estimated measures for the ratees (items) and the residuals for a set of judge-item interactions.

2.2.2 Indices for Detecting Rater Effects

This study focuses on MFRM indices for detecting rater effects. As 'classical' or raw score statistics will be used for the sake of comparison, these are also introduced below.

Leniency v. Severity

Leniency and severity effects manifest as scores which are either lower or higher than those of other raters. MFRM indices rely primarily on the estimated severity measures for the judges and on the separation statistics. For detecting whether individual judges were lenient or severe in relation to the group, a number of indicators are available.

1. Mean scores. Directly comparing the mean scores of the ratings assigned by each judge is the standard indicator within a raw score framework.

Within the MFRM framework, a number of further indicators exist.

2. Judge severity measures. Leniency and severity can be examined directly by comparing the values for the different judges on the judge severity parameter, λ_r. (In an Angoff standard setting, where the only 'examinee' is the BPS as imagined by the different judges, the severity parameter can be omitted and judge severity would appear as different values for the β_n parameter - representing the location of the cut score on the latent variable.)

3. Fixed chi-square test of the hypothesis that the judges share the same level of severity. A significant difference would indicate that *at least two judges differed in severity*.

4. Follow-up *t*-tests. Significant findings on the above chi-square test can be followed up with *t*-tests between pairs of judges, using the judge severity measures and associated standard errors to determine whether the two judges differ significantly in their displayed levels of severity.

5. Judge separation ratio. This ratio measures the spread of the measures for the different judges relative to their precision.

6. Judge separation index. This index indicates the number of statistically distinct severity levels among the raters.

7. Reliability of the judge separation index. This measures the reliability with which the judges have been separated. A value of 0.0 would indicate that the panelists were exchangeable, while higher values indicate that the judges were reliably separated in terms of their severity.

There are no agreed-upon criteria for the above indices. Their value lies in providing information about the degree to which rater severity levels diverged. Actual interpretation remains largely a matter of judgment. In a standard setting, 'interchangeability' of judges is not normally expected. While all of the judges are subject-area experts, they are also chosen to represent diverse backgrounds and may be expected to come to different but defensible interpretations of the performance level descriptors which articulate the standard. It thus becomes a question of judgment on the part of those evaluating the judges' performance as to how much difference is acceptable. Myford & Wolfe (2004b) suggest using *t*-tests to identify judges whose measures differ significantly from one another. Wolfe (2004) flags raters who differ significantly from the group mean. Given the above consideration concerning standard setting judges, another approach would be to define 'problem judges' as those who are at least 2 standard errors (SEs) in distance from any members of the main cluster of judges.

For leniency/severity, there are no clear indicators of group-level effects, which would indicate when most or all of the members of the group were displaying leniency or severity. The only indicators available to detect group-level leniency/ severity are the group level category usage statistics. The problem with attempting to use these to identify group-level effects in an Angoff standard-setting is that it presupposes some prior expectation concerning which categories should be used. If such information existed, standard setting would be unnecessary.

Inaccuracy

Inaccurate ratings are typically diagnosed through correlations and patterns in statistical indicators that are based on residuals.

Two raw-score indices are used here.

1. Raw-score correlations. When scores from an external framework are available, as is often the case in operational Angoff standard settings, these are used. The critical value of the correlation coefficient can be used to flag problematic raters.

2. Single Rater/Rest of Rater (SR/ROR) Correlations. Inter-rater correlations are often used when no external scores are available. The Facets software package calculates this raw score statistic. The critical value of the correlation coefficient can be used to flag problematic raters.

In addition to these raw score statistics, four MFRM indices have been proposed for use in investigating individual level inaccuracy effects.

3. Point-measure correlation. This is the correlation between scores assigned to a group of ratees (items) by a particular examinee and the Rasch parameter estimates or measures for the same ratees. Low consistency between these two scores should be reflected in a low correlation. The critical value of the correlation coefficient can be used to flag problematic raters.

4. Score-expected correlations. The Facets software program generates an expected score for each rater-ratee interaction. A low correlation between observed and expected scores would indicate inaccuracy. The critical value of the correlation coefficient can be used to flag problematic raters.

5. Standard deviation of the residuals. Accurate ratings would result in small, randomly distributed residuals. A large standard deviation of the residuals would thus indicate inaccuracy. Wolfe (2004) 'arbitrarily' defined large as 1.25, and small at .75.

6. Judge fit statistics. These should be sensitive to rater inaccuracy. For mean square fit indicators, values above 1.4 are typically used to flag raters for misfit. For standardized fit statistics, values above 2.0 are used.

Indices have also been proposed for detecting group-level effects.

7. Item separation statistics. Myford & Wolfe (2004) suggested examining the item separation statistics for the ratees (fixed chi-square test of the hypothesis that items share the same measure with a non-significant chi-square value suggesting a group-level effect, item separation ratio with a low ratio suggesting a group-level effect, item separation index with a low value suggesting a group-level effect and reliability of the item separation index with a low value suggesting a group-level effect) for evidence that the raters or judges did not effectively discriminate between or 'separate' the ratees.

Centrality/Extremism

Centrality results in ratings regressing towards the perceived mean of the stimuli range, while extremism results in ratings that cluster near the extremes of the distribution. For detecting individual-level rater centrality and extremism effects, the following indices have been proposed.

One raw-score indices is commonly used in detecting centrality/extremism effects.

1. Standard deviation. With centrality, because the observed ratings form a narrower or more tightly compressed distribution around the mean of the distribution than do the expected ratings, a standard deviation that is smaller for observed than for expected ratings is an indicator of the presence of the effect. Conversely, since the distribution of ratings around the mean is more dispersed where extremism is present,

this effect is indicated by a standard deviation that is larger for observed than for expected ratings. A weakness of the use of the standard deviation as an indicator is that random error would also be expected to inflate the standard deviation, making it difficult to distinguish between accuracy and centrality, on the one hand, and inaccuracy and extremism on the other (Wolfe, 2004; Yue, 2011). Nonetheless, in a simulation study, Yue (2011) found the standard deviation to be one of the better indices for detecting centrality.

Additionally, a number of latent trait indices have been suggested.

3. Standard deviation of the residuals. Centrality would likely result in relatively small residuals, whereas raters displaying extremity will produce residuals with large standard deviations. Wolfe (2004) 'arbitrarily' defined large as 1.25, and small at .75.

4. Judge fit statistics. Although widely used, there is considerable ambiguity concerning how fit statistics might respond to rater centrality/extremism. Research has indicated that centrality may not manifest consistently in fit indices (Wolfe et al., 2000) and Myford & Wolfe (2004) argue that centrality might manifest in fit statistics that were either too low *or* too high. In her simulation study, Yue (2011) found that fit was not an effective indicator of centrality.

5. Expected-residual correlation ($r_{exp,res}$). Proposed by Wolfe (2004), this is the correlation between model-predicted ratings and the residual (observed minus expected rating) for each rater-ratee interaction. Negative correlations indicate centrality, and positive correlations indicate extremism. This is so because raters displaying centrality would assign higher than expected ratings to ratees with low expected values, resulting in positive residuals, *and* lower than expected ratings to ratees with high expected values, resulting in negative residuals. Raters displaying extremism would show precisely the opposite pattern (negative residuals for ratees with low expected values and positive residuals for ratees with high expected values), resulting in a positive correlation. In a simulation study, Yue (2011) found this to be

an effective indicator of centrality. The critical value of the correlation coefficient can be used to flag problematic raters.

6. Measure-residuals correlation ($r_{measure,res}$). The relationship between the ratee parameter estimates or measures would be expected to show the same patterns, for the same reason, since ratees with low measures would have low expected ratings or scores, and vice-versa. Thus, again, negative correlations would indicate centrality and positive correlations would indicate extremism. A simulation study by Yue (2011) found this to be an effective indicator of centrality. The critical value of the correlation coefficient can be used to flag problematic raters.

Indices have also been proposed for detecting group-level effects.

7. Item Separation statistics. Myford & Wolfe (2004) suggested examining separation statistics for the item (fixed chi-square test of the hypothesis that item share the same measure, item separation ratio, item separation index and reliability of the item 0 separation index) for evidence that the raters or judges did not effectively discriminate between or 'separate' the ratees.

8. Item fit statistics. Myford & Wolfe (2004) also suggest that low (overfitting) infit mean square statistics would indicate less variability than expected in the group estimates for the *traits*, where different *traits* are being evaluated. As centrality could be a cause of this lack of discrimination, low fit statistics for the *traits* could be an indicator of centrality. If the same reasoning holds for the items in an Angoff setting, overfitting (lower than expected) item infit mean square statistics might indicate group-level centrality.

Raw score and latent trait/MFRM indices for the detection of rater effects are summarized in Table 2.1

Table 2.1. *Summary of Indicators for Detecting Rater Effects*

Indice	Leniency / Severity	Inaccuracy	Centrality/ Extremity
Raw Score			
Mean score	Unexpectedly Low/ High		
P-value correlations		Low	
SR/ROR correlations		Low	
Standard deviation			Lower/higher than expected
MFRM/Latent Trait Individual-level			
Judge Separation Statistics	High / Significant differences		
Point-Measure Correlation		Low	
Score-Expected Correlation		Low	
Judge fit statistics		Low	Low or High
Residual-Expected Correlation			Negative/Positive
Residual-Measure Correlation			Negative/Positive
MFRM/Latent Trait Group-level			
Item Separation Statistics		Non-sig. chi-sq / Low	Non-sig. chi-sq / Low
Item Infit Mn. Sq. Statistics			Low

2.3 Assumption of the Use of Latent Trait Models

2.3.1 Normative versus Criterion-referenced Contexts

Before discussing the assumption these models make in detecting rater effects, it is essential to first make a distinction which is central to this study. This distinction, rarely explicitly made in the rater effects literature, is between rater deviations *from the ratings of a group or community of raters* and rater deviations *from the ideal of error-free measurement*. Following Wolfe (2004), we can consider the former type as relevant to *normative-referenced settings* and the latter type as relevant to *criterion-referenced settings.* In a normative-reference setting, scores "derive their meaning from their relative standing in the distribution of ratings" (Wolfe, 2004, p. 39). Rater deviations are understood as deviations from a *consensus* understanding of ability. In a criterion-referenced setting, examinee performances are being evaluated against an articulated standard and rater deviations represent errors *in relation to that standard*. Since a standard setting situation is clearly a criterion-referenced setting, the focus here will be on detecting rater deviations from the ideal of error-free measurement.

When the MFRM is used to detect rater effects, the frame of reference normally used is the 'internal frame of reference' constructed by the joint ratings of the raters themselves. When only this judge-generated or internal frame of reference is used, the only claim that can be unproblematically made is that the effects occur *as deviations from the ratings of the other members of the group*. As just noted, this is appropriate for 'normative-reference settings.' However, latent trait/MFRM indices have also been proposed for use in standard setting, a criterion-referenced setting (Eckes, 2009; Engelhard, 2007, 2009, 2011; Engelhard & Anderson, 1998; Engelhard & Cramer, 1997; Englehard & Gordon, 2000; Engelhard & Stone, 1998). Indeed, as one of the advantages of the MFRM-based approach, Engelhard and Anderson explicitly state that it "*does not require data from examinees to define indexes of rating quality*" (1998, p. 227, emphasis added; the authors do, however, immediately add that "Estimates of item difficulties based on examinee data can be used to examine and validate" the results - a suggestion taken up in this study, as detailed below).

What is not always explicitly acknowledged in this literature is that to make

the further claim that ratings deviate from *error-free measurement* requires the critical assumption that "rater effects are randomly distributed and are exhibited by only a minority of raters in the pool" (Wolfe, 2004, p. 47). Put differently, "one must assume that the group of raters assigns, on average, scores that are unbiased" (Wolfe & McVay, 2010, p. 6), or that *no group-level rater effects exist*. As Wolfe further emphasizes:

> A very important topic that has received little attention in the literature relating to rater effects is the interpretive frame of reference within which rater effects are portrayed. A serious shortcoming of the methods described in this article is their reliance on an implicit assumption that rater effects are distributed in the pool of raters in a non-systematic manner. (Wolfe, 2004, p. 48)

A key purpose of this study is to investigate how plausible this assumption is in the context of an Angoff standard setting and to assess how robust the model is to violations of this assumption.

2.3.2 Frames of Reference: Internal versus External

The assumptions being evaluated in this study can be stated in terms of different measurement frames of reference. An *internal* frame of reference "depicts the characteristicts of a particular rater in the context of the characteristics of the pool of raters of whom the rater is a member. To create a relative frame of reference, rating data from the pool of raters is scaled, and parameters are jointly estimated for examinees and raters" (Wolfe & McVay, 2010, p. 9). An *external* frame of reference "depicts the characteristics of a particular rater in the context of the characteristics of scores that are external to the pool of raters of whom the rater is a member. These external scores could have been produced by a pool of expert raters, or the scores could be based on the examinees' performance on an external test." To construct an external frame of reference, "rating data from the pool of raters is scaled while fixing the characteristics (i.e., anchoring the parameters) of examinees on measures that are based on external scores" (Wolfe & McVay, 2010, p. 10).

In an Angoff standard setting, the ability level of the just-proficient student set by the judges within the standard setting frame of reference is *assumed* to correspond to the location of such a student within the test frame of reference. The many-facet

Rasch model has been proposed for use in evaluating this assumption. However, the use of these models requires the further assumption that no group-level rater effects exist. This is so because the expected values against which the observed values are compared were generated from *within* the 'internal' frame of reference created by the raters or Angoff judges themselves. The claim that these indices can detect deviations from error-free measurement requires the assumption that the expected values from the internal frame of reference are the same as expected values from an external 'error-free' frame of reference.

An Angoff standard setting differs from most rating situations in that *an external frame of reference often does exist*. The test items are scaled within both frames of reference: the external frame of reference resulting from the administration of the original test, and the internal frame of reference constructed from the judge's estimates for each of the items. As the task of Angoff judges is to consider how a just-proficient student from the same student population as the actual examinees would perform on the items, it seems reasonable to use results from the administration of the exam to construct the external frame of reference. In fact, in the form of correlations with empirical p-values (Cizek & Bunch, 2007) or with such indices as van der Linden's consistency index (van der Linden, 1982), results from the test frame of reference are commonly used to evaluate the performance of Angoff judges. In this study, results from the external frame are additionally used to examine the central assumption underlying use of the MFRM for detecting rater effects.

CHAPTER THREE METHODS

3.1 Methodological Overview

The data for this study was drawn from two sources. Item parameter estimates were taken from a set of equated English proficiency examinations administered to university students at a private university in Taiwan. These estimates were used in the construction of the external or 'test' frame of reference. Additionally, judge estimates from the first round of a modified Angoff standard setting meeting held to link these exams to the Common European Framework of Reference were used. These estimates were used in the construction of an internal or 'Angoff judge' frame of reference. This data was used in an attempt to answer this study's two primary research questions.

The first question was whether the assumptions underlying use of latent trait indices for the detection of rater effects were likely to be reasonable in a typical modified Angoff standard setting situation. To answer this question, a latent-trait based analysis was conducted twice, using first an internal, then an external frame of reference. In the first, internal analysis, the expected values used to generate the indices were drawn solely from the estimates of the Angoff judges. This situation characterizes most rating situations, where only the ratings made by the raters is available in the construction of the frame of reference. A second analysis was then conducted using the ratings of the Angoff judges *but with the item difficulty measures anchored to their values from the original administration of the exams (i.e., their item bank values).* Applications of the MFRM for detecting rater effects in an Angoff setting *assume* that the results of these two analyses will be the same. The first part of the study directly evaluates this critical assumption.

The second question concerned the assumptions of the Angoff method. The presence of rater effects would represent *violations of these assumptions*. Therefore, in the second part of this study, both classical test theory ('raw score') statistics and MFRM statistics are used attempt to determine whether rater effects were present in the standard setting.

3.2 Exam Items and Calibrations

The exam is the university's "English Proficiency Test" ("EPT"), designed to measure English listening and reading skills in personal, public, educational and occupational contexts. It is administered annually to undergraduate non-English majors at the university during the spring midterm examination period. The composition of the EPT is detailed in Table 3.1.

Table 3.1. *Contents of the English Proficiency Test (EPT)*

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

Items for the EPT are drawn from the EPT item bank. All items are calibrated onto a single scale using the Rasch model. This standard setting project was designed to establish cut scores along the score 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 a composite form, with the constituent items drawn from the different test forms administered during the midterm examination period for first, second, third and fourth year students. The tests shared a number of items which were used to equate them and calibrate them onto a common scale.

As noted, the test items used for this exam were drawn from test forms administered as part of the annual EPT for all four year levels of the program. Items were chosen to represent the same construct as the administered EPT, and on the basis of item quality as assessed by point-biserial correlations and Rasch fit indices (item quality statistics are presented in Appendix A). Items were placed on the composite forms based on their locations in the original exams from which they were drawn. Table 3.2 details the content of the two composite forms.

Table 3.2. *Items on Test Forms Used in the Angoff Standard Setting*

Skill	Items	Description
Listening	01 - 16	What's next?
	17 - 28	Dialogues
	29 - 40	Extended Listening
Reading	01 - 10	Fill in the Blank
	11 - 26	Cloze Reading
	27 - 40	Reading with Questions

3.3 Angoff Standard Setting

In the Angoff study, 18 judges were asked to set a cut score for the B1 level of listening and reading proficiency on the Common European Framework of Reference proficiency scale. A one-day training/orientation session was held on Saturday, July 10, 2010, and three separate Angoff meetings were held on Monday, July 12, Wednesday, July 14, and Friday, July 16. The participants and procedures are described below.

3.3.1 Participants

The 18 Angoff judges were chosen for their experience in language learning/teaching and assessment in a Taiwanese university setting. Amongst judges with such experience, diverse perspectives were sought. Thus, an attempt was made to include native and non-native English speaking teachers at the university, individuals with both EFL teaching experience and administrative responsibilities within the university, teachers from outside institutions and recently graduated students currently

employed in the EFL field. There were 18 judges in all, all of whom received training on the same day. Then, they were broken into sub-panels with six members each to perform the actual standard setting. The composition of the three Angoff panels is presented in Table 3.3.

Table 3.3. *Angoff Judges*

Group	Judge	Gender	English NS/NNS*	Position
I (Mon)	J01	F	NNS	Administrator, former teacher
	J02	F	NNS	Teacher
	J03	M	NNS	Teacher
	J04	F	NS	Teacher
	J05	F	NNS	Teacher
	J06	F	NNS	Teaching Assistant, recently graduated student
II (Wed)	J07	M	NS	Teacher
	J08	M	NS	Teacher, External University
	J09	F	NNS	Teacher
	J10	M	NNS	Teacher
	J11	F	NNS	Teaching Assistant; recently graduated student
	J12	M	NS	Teacher
III (Fri)	J13	F	NNS	Administrator, Teacher
	J14	F	NNS	Administrator, former teacher
	J15	F	NNS	Teacher
	J16	F	NNS	Teacher, External University
	J17	F	NNS	Teacher
	J18	F	NNS	Teacher

*NS = 'native speaker'; NNS = 'non-native speaker'.

3.3.2 Procedures

A group of six judges participated on each day. (The procedure was conducted on three separate days to ensure that proper procedures were followed, particularly during the dicussion period. In some implementations of the Angoff procedure, larger groups are accommodated by having 'table leaders' who are themselves participants take charge of the discussion. Given the absence of experienced table leaders, it was decided that it would be preferable to have the facilitators present for each discussion session. This required having the groups meet on separate days.)

Five days prior to the training session, an email was sent to all judges containing an introductory letter, an agenda for the training session, the training materials, and two forms collecting personal information and agreements concerning test security and informed consent for the research portion of the project. The introductory letter contained a link to a CEFR familiarization website, <www.CEFtrain.net>. The training materials consisted of pages 33-36 from the CEFR (in slightly adapted form), which summarize the CEFR levels; the listening and reading components of the CEFR self-assessment grid (CEFR Table 2); and a link to the website. 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. These homework tasks were developed after referring to page 18 of the CEFR linking manual (Council of Europe, 2009).

On the day of the 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. They then commenced the CEFR familiarization process. After a brief description of the CEFR, they were given a sheet containing the global level 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.

The session then shifted to the performance level descriptors (PLDs) for the CEFR reading scale. (The CEFR scales used for the development of both the reading and listening PLDs are listed in Appendix B, along with the global descriptors which provide a general idea of the content of the different levels.) The first activity was another sorting activity, in which judges were asked to sort (individually, and then in pairs) 20 CEFR reading descriptors from CEFR levels A1 to B2. Then they were given the CEFR reading descriptors from the scales used in the study, for CEFR levels A1 to B2. Next, judges were given a 13-item reading test, taken from training material made available by the Council of Europe (CoE). For each item, the judges were asked

to first attempt to answer the item and then to assign the item to a CEFR level, based on the skills required to correctly answer the item. After sharing their answers in pairs, the answers from the original CoE study were shown and discussed. This concluded training for the reading PLDs.

The training for the listening PLDs was conducted in parallel fashion. Judges were asked, individually and in pairs, to sort 20 PLDs 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 for levels A1 to B2. Judges were then given a 6-item listening test, taken from the CoE training material mentioned above, and asked to attempt to respond to the item and then assign a CEFR-level to the item based on its perceived difficulty. Judges shared their answers in pairs, and then the recommended answers from the CoE study were shown and discussed by the whole group. This concluded training for the listening PLDs.

After a break for lunch, the judges took both tests. The purpose of this was to enhance their ability to see what makes items difficult by seeing them, to some extent, from a student's perspective. The judges were then asked to sit together in a circle with the other members of their standard setting panel. A group leader was chosen, and each group 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. 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. They were asked to refer to their listening and reading PLDs for the A2 and B1 levels, and to summarize on the forms what they felt were 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 then told that when they returned for the actual meeting, they would have one training round prior to the meeting and would then perform the actual standard setting. This concluded the training session.

The Angoff meetings were held on July 12, 14 and 16, 2010. At each meeting,

standards were 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 notion of the B1 BPS. They were then asked to estimate, based on their knowledge of students in the university's English program (or Taiwanese university students in general for judges from other universities), the percentage of students who had reached the B1 level for the skill in question. They were asked to write this estimate down. Then, the test form and the round 1 rating form for the reading test were distributed to the judges. 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 (see Appendix C). 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. Judges were then given a practice round, in which they were asked to pencil in their ratings for the first couple of items. It was made clear that this was simply a practice round, designed 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 conducted. Two further rounds were conducted. Between each round, the judges were shown empirical p-values from the actual administration of the test. Between the second and third rounds, judges were also shown the number of students who would reach B1 based on the cut score they (as individuals) had set in the second round. Judges were also asked to share and discuss their ratings between rounds. Since the purpose of the present study is to assess whether the assumptions of the method hold in the *absence* of such feedback data, only the first round of results is used for the main analyses conducted below.

The same procedures were followed for the listening test. (Since the test form did not contain the scripts for the listening passages, a separate form was created for the listening test which contained both the listening scripts and the associated items for the judges to refer to during the actual standard setting meeting.)

3.4 Analysis

Scoring data were scaled to a Rasch rating scale model using the Facets software program (Linacre, 2009). For both the reading and listening exam, scaling and parameter estimation was carried out twice, to estimate parameters for two distinct frames of reference: the 'internal' frame of reference constructed by the Angoff judges, and the 'external' frame of reference using the item difficulty information from the administered exams. The binomial model was used to estimate parameters (Eckes, 2009; Engelhard & Anderson, 1998; J.M. Linacre, personal communication, July 7, 2010). Following Engelhard & Anderson (1998), the 101 possible category levels corresponding to probability estimates were collapsed into an 11-point scale by recoding the counts obtained from the judges as follows:

0 - 5 = 0
6 - 15 = 1
16 - 25 = 2
26 - 35 = 3
36 - 45 = 4
46 - 55 = 5
56 - 65 = 6
66 - 75 = 7
76 - 85 = 8
86 - 95 = 9
96 - 100 = 10

Although some information is lost through recoding, the empty cells and small numbers of counts in many other cells makes MFRM analysis difficult.

The analysis was conducted in two phases. The first phase tested the assumption of the use of the MFRM for detecting rater effects by comparing results from the internal and external frameworks. The second phase used the indices to test the assumptions of the modified Angoff method.

3.4.1 Evaluating the Assumption of the MFRM

In the first phase, parameters were estimated for the internal and external frameworks. For both frameworks, values for the latent trait indices used to detect rater effects were generated for each rater. Results from the internal and external frameworks were

compared, to determine their level of agreement. As a further check, 'classical' or 'raw score' indices for detecting rater effects were also included and used to assess the performance of the different latent trait indices. The reading test was analyzed first, and the listening test after it, to see whether the results were replicated. Details for each rater effect are as follows.

Leniency/Severity

No direct measures of judge severity (i.e., the cut score) are available in the external framework, so only indirect comparison is possible. This was performed using separation statistics. As noted earlier, there is no clear guidance on flagging judges. Since some degree of diversity is expected, and indeed encouraged, in a standard setting, in this study, judges were flagged only if their severity measures (cut scores) differed significantly from all other judges in the main group. Leniency/severity indices and the criteria used to flag raters are listed in Table 3.4.

Table 3.4 *Leniency/Severity - Indices and Criteria*

Indice	Flagging Criteria for Leniency / Severity
Raw Score	
Mean score	Unexpectedly Low/High
MFRM/Latent Trait Individual-level	
Fixed Chi-square	No set criterion - Significant results warrant investigation
t-tests between judge pairs	Judges significantly different from the nearest judge in the main group of judges flagged
Judge Separation Ratio	No set criterion - values above 1 warrant investigation
Judge Separation Index (Strata)	No set criterion - values above 1 warrant investigation
Reliability of Separation Index	No set criterion; values above 0 warrant investigation

Inaccuracy

Correlations and fit statistics have both been proposed for use in detecting inaccuracy. Two raw score indices were used. As is standard in Angoff settings, correlations between rater estimates and the p-values from the original administration of the test were computed. The Facets software also computes 'point-biserial' or single rater/rest of rater correlations. For latent trait indicators, fit statistics, point-measure correlations and score-expected correlations were used. Item separation statistics were used to detect group-level effects. Inaccuracy indices and the criteria used are listed in Table 3.5.

Table 3.5 *Inaccuracy - Indices and Criteria*

Indice	Inaccuracy
Raw Score	
P-value correlations	Correlation coefficient below the .05 (2-tailed) level of significance (Above .05 but below .01 flagged as 'marginal')
SR/ROR correlations	Comparatively low
MFRM/Latent Trait Individual-level	
Judge mean square fit statistics	Above 1.4 (mean square) Above 2.0 (standardized)
Point-Measure Correlation*	Correlation coefficient below the .05 (2-tailed) level of significance (Above .05 but below .01 flagged as 'marginal')
Score-Expected Correlation	Correlation coefficient below the .05 (2-tailed) level of significance (Above .05 but below .01 flagged as 'marginal')
MFRM/Latent Trait Group-level	
Item Separation Statistics	Non-sig. chi-sq / Low

*Where Rasch measures are used, the sign is reversed when correlations are calculated for ease of interpretation.

Centrality/Extremism

The only raw score index used was the standard deviation of the estimates made by each judge. Latent trait indices included the standard deviation of the residuals, judge fit statistics, the expected-residual correlation and measure-residual correlations. At the group level, MFRM item separation statistics and item fit indices were used. The indices and the criteria used are listed in Table 3.6.

Table 3.6 *Centrality/Extremism - Indices and Criteria*

Indice	Centrality/ Extremity
Raw Score	
Standard deviation	Lower/higher than expected
MFRM/Latent Trait Individual-level	
SD of the residuals	Below 0.75 / Above 1.25
Judge Fit Statistics	Below 0.6 / Above 1.4 (mean square) Below -2.0 / Above 2.0 (standardized)
Expected-Residual Correlation	Negative / Positive ($p < .05$)
Measure-Residual Correlation*	Negative / Positive ($p < .05$)
MFRM/Latent Trait Group-level	
Item Separation Statistics	Non-sig. chi-sq / Low
Item InFit Mean Square Statistics	Below 0.6 (mean square)

*Where Rasch measures are used in this study, the sign is reversed when correlations are calculated for ease of interpretation.

3.4.2 Evaluating the Assumptions of the Angoff Method

The second phase of the analysis makes use of the results from the first. The most suitable frame of reference (internal or external) as determined in the first phase was used in the second phase. Decisions about the presence or absence of rater effects were made after consideration of all of the available evidence.

CHAPTER 4 RESULTS

4.1 Assumption of the MFRM

4.1.1 Leniency/Severity

No direct measures of the B1 cut score are available. Therefore, for leniency/severity, internal and external results can only be compared *indirectly* through separation statistics generated in both frames.

Reading

The judge separation statistics appear in Table 4.1 and the individual-level raw score and latent trait statistics for individual judges appear in Table 4.2. In Table 4.1, the chi-square value of 292.8 with 17 degrees of freedom is statistically significant ($p < .05$), indicating that the judges did *not* all display the same degree of difficulty - there were significant differences in their B1 cut scores. The judge separation ratio of 3.95 indicates that the differences between the judges' cut scores were about 4 times greater than the error with which they were measured. The judge separation index of 5.60 suggests the presence of five to six statistically distinguishable levels of judge severity. Finally, the reliability index of 0.94 suggests that the judges were reliably separated. Taken together, it seems clear that unwanted variation was present.

In the external 'test' frame of reference, the conclusions are nearly the same, but the values differ. The chi-square value (419.8), separation ratio (4.77), number of distinct strata (6.69) and reliability of the separation index (0.96) all indicate an even *higher degree of dispersion* of cut scores than did results from the internal frame. That is, results from the external frame indicated *less* agreement between judges.

Table 4.1. *Judge Separation Statistics for Reading, Internal v. External Frames*

	READING	
	INTERNAL	**EXTERNAL**
Fixed Chi Square Test	292.8, df = 17, *p* = .00	419.8, df = 17, p = .00
Separation Ratio	3.95	4.77
Separation Index (Strata)	5.60	6.69
Reliability of the Separation Index	0.94	0.96

Given that significant variation exists, the next step is to determine whether individual judges can be flagged for leniency/severity effects. The judge severity estimates (i.e., B1 cut scores), standard errors, and distances from the mean estimate in SE units appear in Table 4.2, where judges are ranked from most lenient to most severe by their severity measures (cut scores). From the table, we can see that the same four judges are flagged for leniency/severity effects in both the internal and external analysis. J11 and J07 are both two or more SEs below the main group of judges and are thus flagged for leniency, and J01 and J08 are both two or more SEs above the main group and are flagged for severity.

Table 4.2. *Reading Results (Means, Severity Measures and SEs), Internal v. External*

	Raw Score	Latent Trait INTERNAL			Latent Trait EXTERNAL		
Judges	Mean (% Correct)	β	SE	Distance to Mean (SEs)	β	SE	Distance to Mean (SEs)
J11	44.5	-0.48*	0.11	-10.30	-0.7*	0.13	-12.48
J07	53.9	-0.09*	0.11	-6.82	-0.13*	0.13	-8.20
J17	62.8	0.32	0.11	-3.17	0.46	0.13	-3.78
J15	63.5	0.37	0.11	-2.72	0.53	0.13	-3.25
J18	64.8	0.43	0.11	-2.19	0.61	0.13	-2.65
J05	69.6	0.56	0.11	-1.03	0.8	0.13	-1.23
J10	70.4	0.63	0.11	-0.41	0.9	0.13	-0.48
J16	69.3	0.66	0.11	-0.14	0.95	0.13	-0.10
J02	71.3	0.71	0.11	0.31	1.02	0.13	0.42
J09	71	0.75	0.11	0.66	1.08	0.13	0.87
J12	71.6	0.78	0.11	0.93	1.11	0.13	1.10
J06	72.4	0.79	0.11	1.02	1.13	0.13	1.25
J14	73.5	0.88	0.11	1.82	1.25	0.13	2.15
J04	76.8	0.97	0.11	2.62	1.38	0.14	3.12
J13	75.6	0.98	0.11	2.71	1.4	0.14	3.27
J03	77	1.05	0.12	3.34	1.5	0.14	4.02
J08	81.4	1.3*	0.12	5.56	1.85*	0.14	6.65
J01	85.3	1.55*	0.13	7.79	2.21*	0.15	9.35
Group Avg.	69.71	0.68	0.11	0.00	0.96	0.13	0

*More than 2 group average SEs apart from next closest judge.

Figure 4.1 provides a visual indication of the differences between the individual judges. Using the error bars, we can see that two judges at either extreme are significantly distinct from the main body of judges.

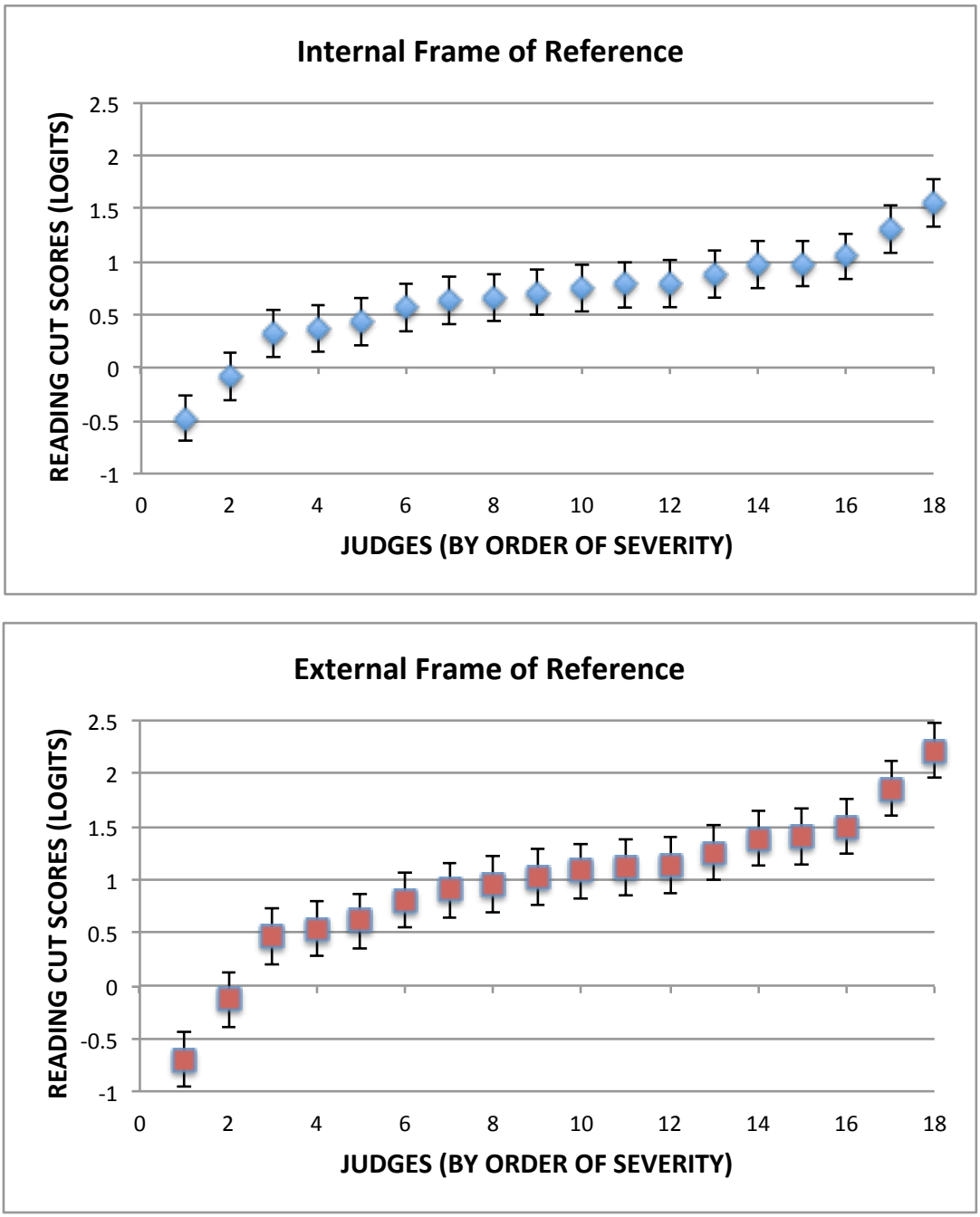


Figure 4.1. Reading cut scores (judge severity measures in logits), internal and external frameworks. Judges are ordered with the most lenient judge on the left and the most severe on the right. Error bars are two group-average SEs in length.

Putting the measures from both frameworks onto a single scale in Figure 4.2 clearly shows that the scale in the internal framework is compressed relative to that for the external framework. Notice that the external range *starts at a lower point* on the scale but *ends at a higher point*. This apparent compression is consistent with the separation statistics in Table 4.1.

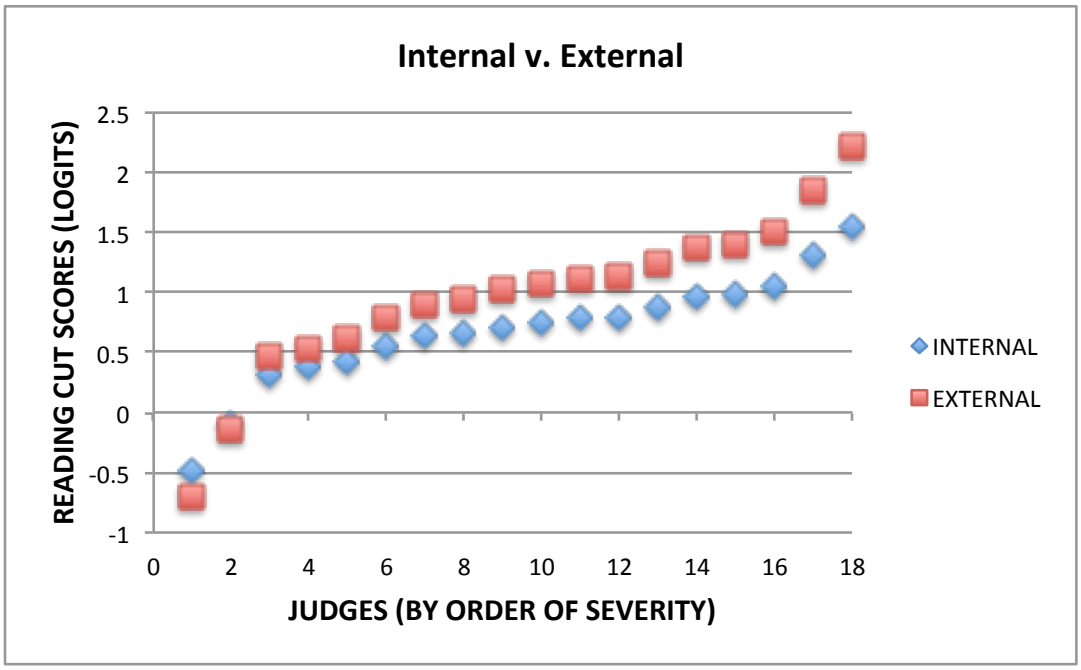


Figure 4.2. Comparison of reading cut scores (judge severity measures in logits), internal and external frameworks. Judges are ordered with the most lenient judge on the left and the most severe on the right.

Listening

In the left-hand column of Table 4.3, the chi-square value of 175.3 with 17 degrees of freedom is statistically significant ($p < .05$), indicating that the judges did *not* all display the same degree of difficulty - there were significant differences in their B1 cut scores. The judge separation ratio of 2.85 indicates that the differences between the judges' cut scores were about 3 times greater than the error with which they were measured. The judge separation index of 4.14 suggests the presence of about four statistically distinguishable levels of judge severity. Finally, the reliability index of 0.89 suggests that the judges have been reliably separated. Taken together, it again seems clear that unwanted variation was present.

Turning to the external frame, we can see that the pattern of differences between the two frames found for reading is replicated for listening. Results for listening differ in the same way as they did for reading. The chi-square value (228.9), separation ratio (3.32), number of distinct strata (4.75) and reliability of the separation index (0.92) all indicate unwanted variation, as in the internal frame. At the same time, the external frame again indicates a *higher degree of dispersion* of rater judgments.

Table 4.3. *Judge Separation Statistics for Listening, Internal v. External Frames*

	LISTENING	
	INTERNAL	**EXTERNAL**
Fixed Chi Square Test	175.3, df = 17, p = .00	228.9, df = 17, p = .00
Separation Ratio	2.85	3.32
Separation Index (Strata)	4.14	4.75
Reliability of the Separation Index	0.89	0.92

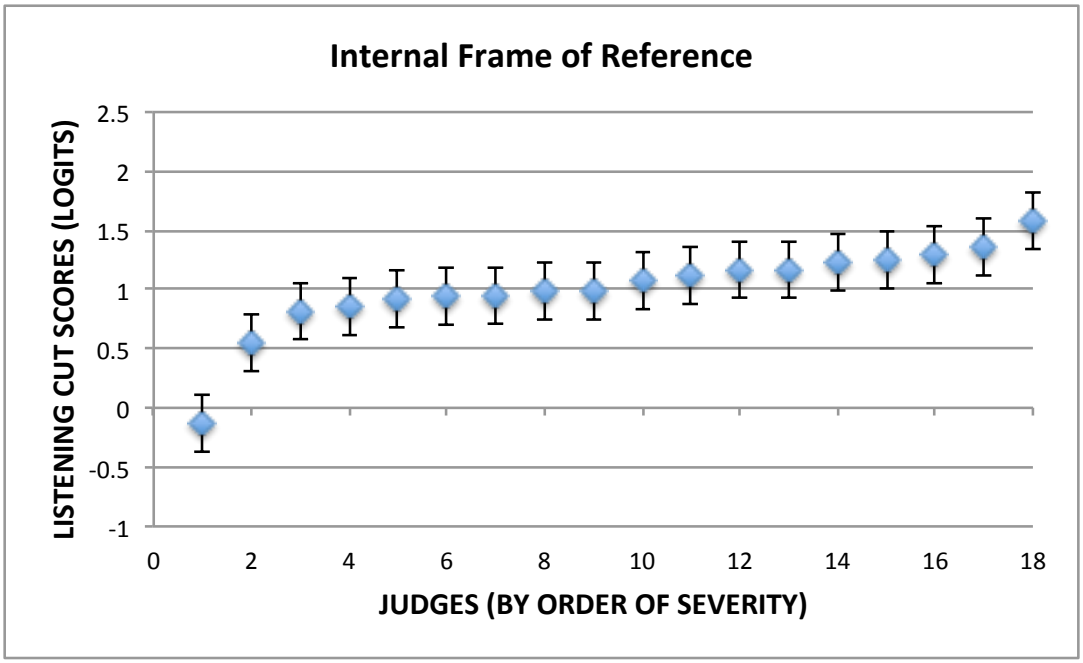
Turning to differences between individual judges in Table 4.4, we can see that in both frames, J10 and J07 are flagged for leniency, based on the criterion of being two or more SEs away from the main group of judges. In the internal frame, no other judges are flagged. *In the external frame, however, J03 is flagged for severity.* Thus, for the listening exam, we again find the higher degree of dispersion that was seen for the reading exam, but in this case, the higher degree of dispersion led to a difference in flagging decisions.

Table 4.4. *Listening Results (Means, Severity Measures and SEs), Internal v. External*

	Raw Score	Latent Trait INTERNAL			Latent Trait EXTERNAL		
Judges	Mean	β	SE	Distance to Mean (SEs)	β	SE	Distance to Mean (SEs)
J10	53.6	-0.13*	0.11	-9.58	-0.17*	0.13	-11.00
J07	66.9	0.55*	0.11	-3.83	0.73*	0.13	-4.33
J09	71.8	0.82	0.11	-1.55	1.08	0.13	-1.74
J11	74.2	0.85	0.11	-1.30	1.11	0.13	-1.51
J12	73.9	0.91	0.12	-0.79	1.2	0.13	-0.85
J06	76.5	0.94	0.12	-0.54	1.23	0.13	-0.63
J17	76.6	0.95	0.12	-0.45	1.25	0.13	-0.48
J13	76.1	0.98	0.12	-0.20	1.28	0.13	-0.26
J15	74.9	0.98	0.12	-0.20	1.28	0.13	-0.26
J04	77.1	1.07	0.12	0.56	1.41	0.13	0.71
J05	79.3	1.11	0.12	0.90	1.46	0.14	1.08
J01	79	1.17	0.12	1.41	1.53	0.14	1.60
J02	78.6	1.17	0.12	1.41	1.53	0.14	1.60
J16	79.8	1.23	0.12	1.92	1.61	0.14	2.19
J14	79.5	1.24	0.12	2.00	1.63	0.14	2.34
J08	81.5	1.29	0.12	2.42	1.68	0.14	2.71
J18	83.9	1.35	0.12	2.93	1.76	0.14	3.30
J03	85.2	1.58	0.13	4.87	2.06*	0.15	5.52
GROUP AVG.	76.02	1.00	0.12	0.00	1.31	0.14	0.00

*More than 2 group average SEs apart from next closest judge.

Figure 4.3 provides a visual indication of the differences between the individual judges. Using the error bars, we can see that two judges at the lower extreme are significantly distinct from the main body of judges, while the most severe judge appears to be a marginal case.



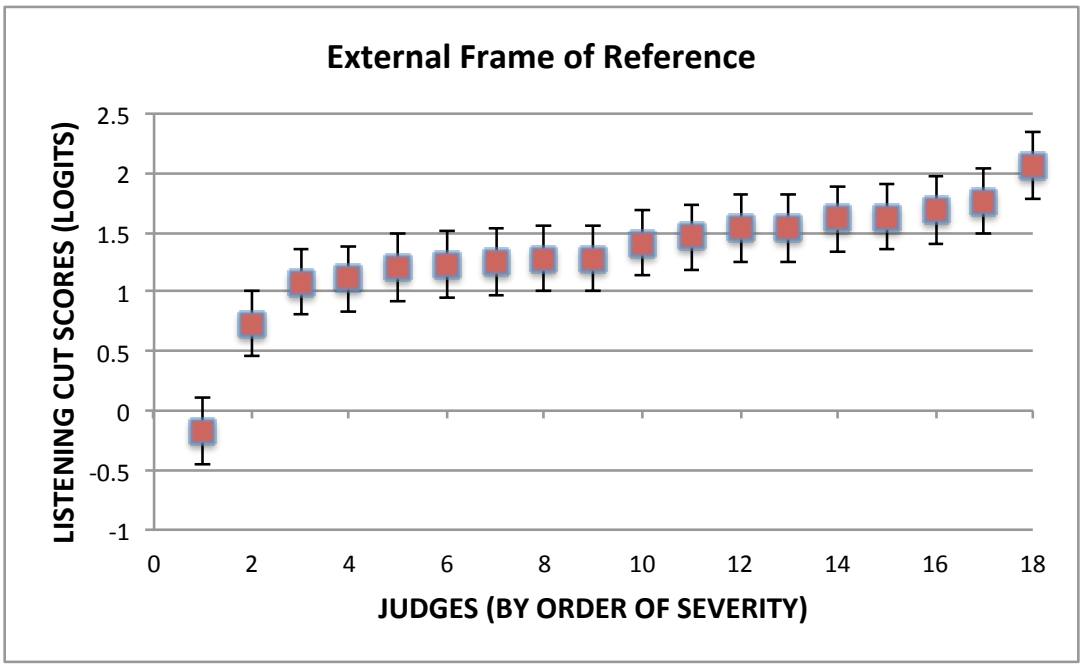


Figure 4.3. Listening cut scores (judge severity measures in logits), internal and external frameworks. Judges are ordered with the most lenient judge on the left and the most severe on the right. Error bars are two group-average SEs in length.

Putting the measures from both frameworks onto a single scale in Figure 4.4 clearly shows that, for listening as for reading, the scale in the internal framework is compressed relative to that for the external framework. Notice that the external range

again *starts at a lower point* on the scale but *ends at a higher point*. This apparent compression is consistent with the separation statistics in Table 4.3.

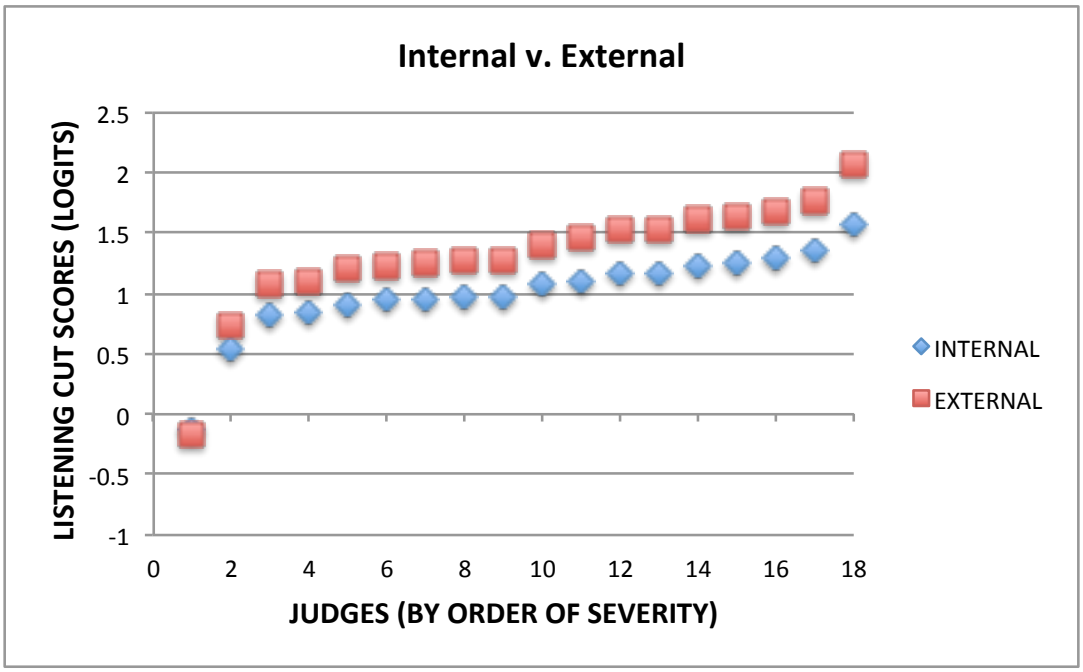


Figure 4.4. Comparison of listening cut scores (judge severity measures in logits), internal and external frameworks. Judges are ordered with the most lenient judge on the left and the most severe on the right.

Summary

Diagnostically, six judges were flagged in the internal frame, but seven judges were flagged in the external frame. For both reading and listening, the scale of judge severity measures was relatively compressed in the internal compared to the external frame, with the result that both the measures and the separation statistics resulted in a relatively more 'optimistic' evaluation of the quality of the ratings. That is, raters appeared to be in greater agreement in the internal than in the external frame. (From the group average measures in Tables 4.2 and 4.4, it can be seen that this scale compression also had the effect of lowering the cut score for both exams in the internal compared to the external frames. This point is returned to below.)

It might be argued that the presence of outlying judges might have somehow skewed the results. To evaluate this argument, the flagged judges were removed and the analysis was performed again. Table 4.5 shows that, even with the flagged judges removed, the scale in the internal frame again appears to be relatively compressed and

the internal frame generated results which were more 'optimistic' in the sense of indicating a greater degree of inter-judge agreement than the results from the external frame. An alternative explanation, based on a group-level centrality effect, will be offered below.

Table 4.5. *Judge Separation Statistics for Reading and Listening, Internal v. External, with Flagged Judges Removed*

	READING		**LISTENING**	
	INTERNAL	**EXTERNAL**	**INTERNAL**	**EXTERNAL**
No. of Judges	14	14	15	15
Fixed Chi Square Test	58.4, df = 13, p = .00	79.5, df = 13, p = .00	27.9, df = 14, p = .01	36.3, df = 14, p = .00
Separation Ratio	1.77	2.15	0.93	1.20
Separation Index (Strata)	2.69	3.20	1.58	1.93
Reliability of the Separation Index	0.76	0.82	0.46	0.59

4.1.2 Inaccuracy

Individual-level Effects

The score-expected and point-measure correlations yielded essentially identical results, as did the residual-based indices (standard deviation of the residuals, fit statistics). Thus, to simplify the presentation of results, only results for the score-expected correlations, infit mean square statistics and the raw score correlations (score/p-value and single-rater/rest-of-rater) are shown below. Results for all indicators can be found in Appendix D. Values for all indices also appear in the correlation matrices below, which indicate the high level of agreement found.

Reading. Values for the selected indices appear in Table 4.6, where raters displaying effects are flagged with asterisks. Starting with the internal framework, the point-biserial (SR/ROR) correlation flagged four judges as marginal (J04, J08, J11, J12). Note that with the default manner of calculating the point-biserial (SR/ROR) correlation using the Facets software, the judge for whom the correlation is being calculated is not included in deriving the values for the group; since judges correlate perfectly with themselves, removing them deflates the statistic, giving a more conservative figure (Linacre, 2009). The score-expected correlation flags three raters as being either marginal (J07, J08) or inaccurate (J10). In the internal framework, no judges are flagged for misfit. In fact, with the exception of J03 and J07, all of the judges actually overfit the model, meaning that their ratings were actually more predictable than expected. Notice that only J07 is flagged by both the correlational and residual-based indices.

Turning to the external framework, the results are very different. The score/p-value correlations and the score-expected correlations agree in flagging six judges for inaccuracy (J06, J07, J08, J10, J11 and J12) and three as marginal (J01, J04 and J17). Using the fit statistics, all judges except for J13 are flagged for misfit. The correlational and residual-based indices both flag far more raters in the external framework. Note also that the residual-based indices have gone from being *less conservative* than the correlational indices in the internal framework to *more conservative* in the external framework.

Table 4.6. *Indices of Inaccuracy for Reading, Internal v. External*

	Correlation of Scores with p-values		Point-Biserial Correlation		Score-Expected Correlation		InFit Mean Square	
Judges	INT.	EXT.	INT.	EXT.	INT.	EXT.	INT.	EXT.
J01		0.33*	0.59		0.64	.32*	0.56	1.87**
J02		0.54	0.79		0.83	0.54	0.39	1.55**
J03		0.57	0.56		0.64	0.56	1.11	1.89**
J04		0.38*	0.38*		0.44	0.37*	0.63	1.77**
J05		0.43	0.44		0.49	0.40	0.42	1.64**
J06		0.21**	0.43		0.50	0.24**	0.70	2.36**
J07		-0.04**	0.25**		0.34*	-0.03**	1.01	3.58**
J08		0.31**	0.35*		0.39*	0.29**	0.56	1.73**
J09		0.43	0.59		0.64	0.44	0.43	1.66**
J10		0.06**	0.24**		0.3**	0.04**	0.55	2.33**
J11		0.12**	0.40*		0.46	0.10**	0.69	2.78**
J12		0.30**	0.39*		0.46	0.29**	0.61	2.03**
J13		0.63	0.61		0.65	0.64	0.35	1.05
J14		0.54	0.76		0.80	0.55	0.57	1.64**
J15		0.41	0.77		0.82	0.43	0.81	2.65**
J16		0.55	0.79		0.83	0.55	0.81	2.07**
J17		0.38*	0.71		0.76	0.39*	0.57	2.28**
J18		0.65	0.81		0.85	0.68	0.78	1.55**
Accurate		9	12		15	9	18	1
*Marginal		3	4		2	3	na	na
**Inaccurate		6	2		1	6	0	17

*Marginal, $p > .01$; **Inaccurate, $p > .05$ for correlations; Fit Mean Square > 1.40. (Note that asterisks for correlations indicate low or non-significant correlations.)

In short, it is clear that results from the internal framework do not correspond to results in the external framework. Figure 4.5 provides a visual illustration of the differences between the frameworks for the score-expected correlations and infit mean square statistics. The score-expected correlation shows that a reasonable linear relationship is maintained across the two but that the correlations are noticeably higher within the internal framework. For the residual-based statistics, the values in

the internal framework appear significantly compressed compared to those in the external framework.

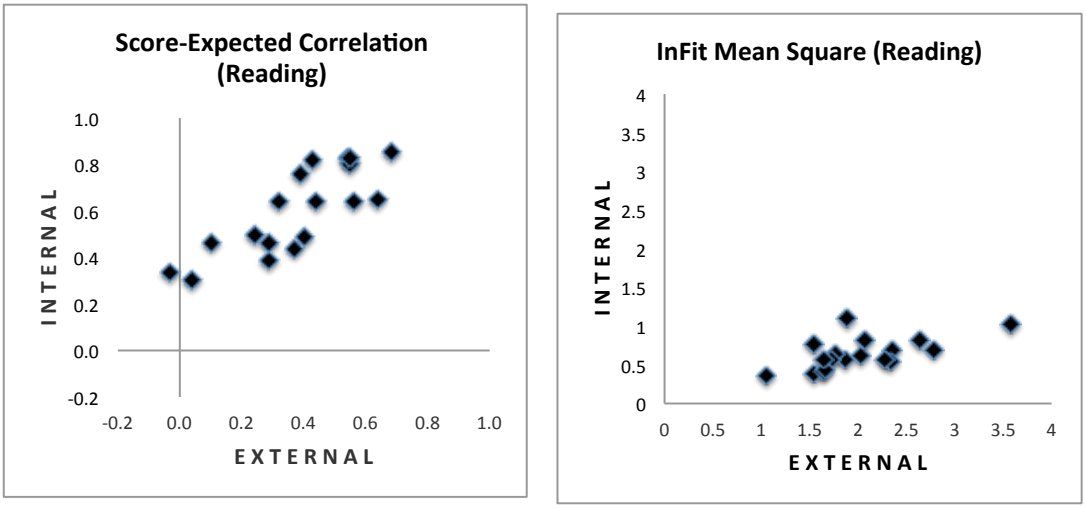


Figure 4.5. Inaccuracy indices for reading, internal v. external frameworks.

Above, it was noted that the correlational and residual-based indices were not in close agreement with each other. The correlation matrices in Table 4.7 allow us to investigate this more carefully. The raw score score/p-value correlations serve as a useful 'standard' for judging how well the other indices measure accuracy. In the internal framework, all of the correlational indices are in good agreement with the score/p-value correlations. The residual-based indices, however, show only a weak relationship with score/p-value correlations, which fails to reach statistical significance. Interestingly, this changes sharply in the external framework. Since the score-expected and point-measure correlations are now based on the same data as the score/p-value correlations, it is not surprising that they show a near-perfect correlation. More interesting is that the fit indices actually correlate reasonably well with the p-value correlations in the external framework. In other words, they were effective at signaling inaccuracy in the external, but not the internal framework.

Table 4.7. *Correlation Matrices of Inaccuracy Indices for Reading*

	Corr P-values	SR/ROR	Score-Expected	Point-Measure	SD Residuals	InFit MnSq	InFit Z	OutFit MnSq
INTERNAL								
SR/ROR	.798**							
Score-Expected	.787**	.998**						
Point-Measure	.777**	.998**	0.990					
SD Residuals	-0.243	-0.058	-0.012	-0.004				
InFit MnSq	-0.180	-0.069	-0.022	-0.016	0.966**			
InFit Z	-0.224	-0.085	-0.040	-0.033	0.968**	.992**		
OutFit MnSq	-0.230	-0.090	-0.042	-0.036	.985**	.992**	.988**	
OutFit Z	-0.260	-0.103	-0.057	-0.050	.983**	.984**	.994*	.993**
EXTERNAL								
SR/ROR	.798**							
Score-Expected	.996**	.822**						
Point-Measure	.992**	.839**	.997**					
SD Residuals	-.741**	-0.345	-.713**	-.727**				
InFit MnSq	-.795**	-.410**	-.768**	-.774**	.981**			
InFit Z	-.787**	-0.391	-.763**	-.767**	.983**	992**		
OutFit MnSq	-.850**	-.515*	-.830**	-0.845**	.973**	.981**	.971**	
OutFit Z	-.847**	-.505*	-.830**	-.843**	.976**	.975**	.980**	.992**

* Significant at 0.05 level (2-tailed). **Significant at 0.01 level (2-tailed).

Figure 4.6 visually displays this relationship between selected indices and the score/p-value correlation. For the score-expected correlations, we see the same reasonably accurate linear relationship but with generally higher (more ‘optimistic’) correlations in the internal framework. For the residual-based indices, it seems as if the compression of the values also suppresses any correlation with judge accuracy. In the external framework, the expected relationship appears: judges with high score/p-value correlations better fit the model and fit worsens as the judges become less accurate.

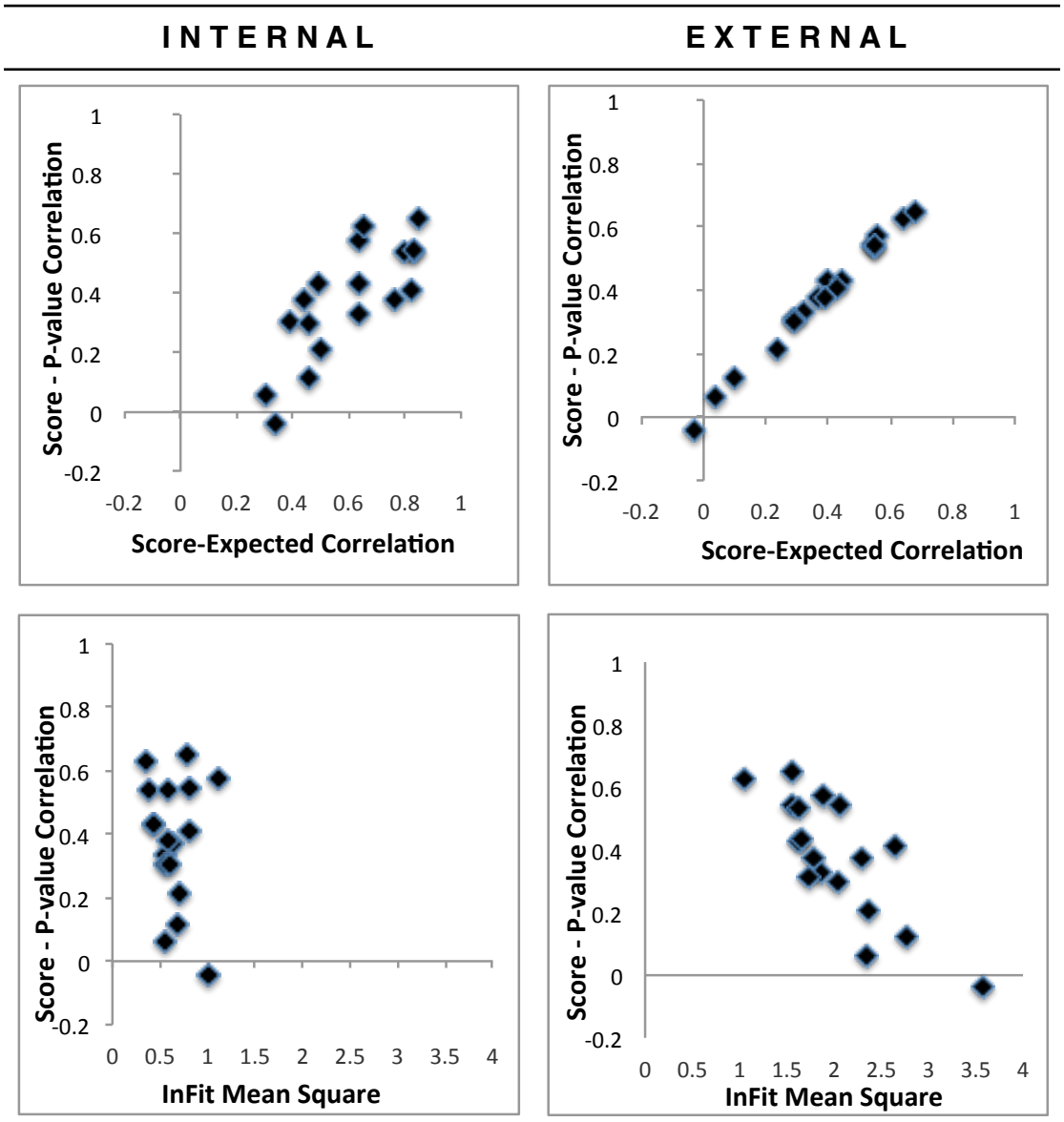


Figure 4.6. Inaccuracy indices v. score/p-value correlations, internal v. external.

Listening. Values for the correlational indices appear in Table 4.8 where raters displaying effects are flagged with asterisks. Starting with the internal framework, the point-biserial (SR/ROR) correlation flags J11 as inaccurate and J07 as marginal. The score-expected correlation flags J11 as inaccurate while all other judges are categorized as accurate. In terms of fit, in the internal framework, all judges actually *overfit* the model (mean square below 1.0).

Turning to the external framework, the results again differ in the same direction as for reading. The score/p-value correlations and score-expected correlations in Table 4.8 agree in flagging J07 and J11 for inaccuracy and J05 and J17

as marginal. J18 was also flagged as marginal by the score-expected correlation and J03 by the score/p-value correlation. Turning to the residual-based indices, J07 and J11 are both flagged for misfit. Note here that there is a higher level of agreement between the different indices. As with reading, more judges were flagged in the external framework, although the difference was less severe for listening.

Table 4.8. *Indices of Inaccuracy for Listening, Internal v. External*

	Correlation of Scores with p-values		Point-Biserial Correlation		Score-Expected Correlation		InFit Mean Square	
Judges	INT.	EXT.	INT.	EXT.	INT.	EXT.	INT.	EXT.
J01		0.45	0.7		0.74	0.48	0.29	1.06
J02		0.72	0.66		0.71	0.71	0.30	0.61
J03		.433*	0.58		0.62	0.48	0.28	0.89
J04		0.56	0.69		0.74	0.56	0.51	1.14
J05		.368*	0.6		0.65	0.39*	0.31	1.20
J06		0.47	0.64		0.68	0.47	0.22	1.02
J07		0.29**	0.34*		0.41	0.23**	0.47	1.57**
J08		0.42	0.7		0.74	0.41	0.25	1.12
J09		0.55	0.71		0.75	0.55	0.28	0.98
J10		0.53	0.48		0.57	0.61	0.57	0.99
J11		-0.06**	0.02**		0.12**	-0.06**	0.89	2.37**
J12		0.69	0.67		0.72	0.67	0.28	0.71
J13		0.66	0.62		0.67	0.68	0.23	0.69
J14		0.57	0.59		0.66	0.56	0.50	1.01
J15		0.65	0.69		0.74	0.67	0.36	0.76
J16		0.60	0.73		0.78	0.60	0.34	0.87
J17		.395*	0.77		0.80	0.37*	0.19	1.25
J18		0.46	0.72		0.76	0.41*	0.20	1.06
Accurate		13	16		17	13	18	16
*Marginal		3	1		0	3	na	na
**Inaccurate		2	1		1	2	0	2

*Marginal, $p > .01$; **Inaccurate, $p > .05$ for correlations; Fit Mean Square > 1.40. (Note that asterisks for correlations indicate *low* or *non-significant* correlations.)

Figure 4.7 shows the same tendencies as appeared for reading. For the score-expected correlations, the linear relationship was good, but correlations were again higher in the internal framework. For infit, again, the values were relatively 'compressed' in the internal framework.

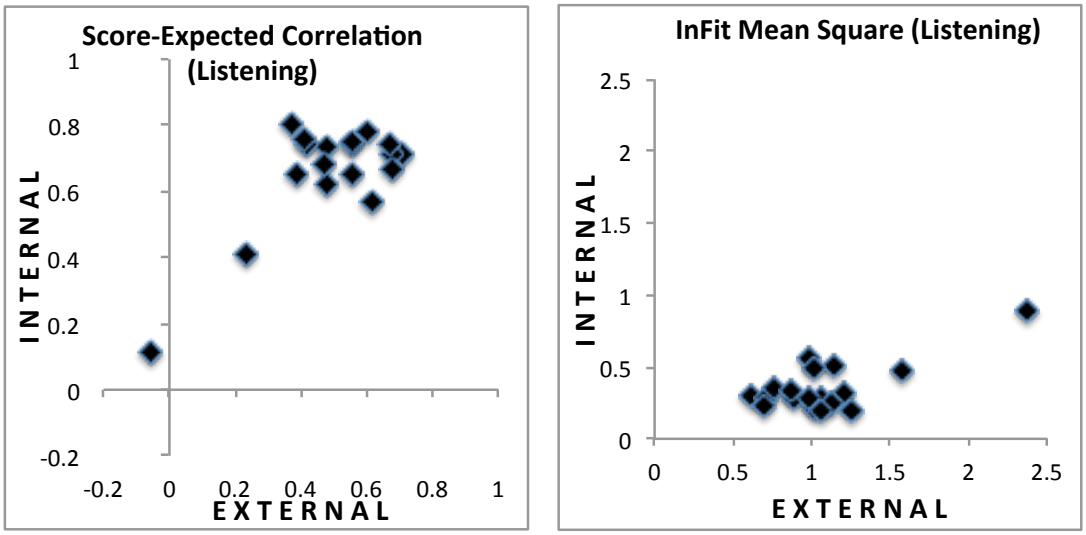


Figure 4.7. Inaccuracy indices for listening, internal v. external frameworks.

The correlation matrices in Table 4.9 allow us to investigate the relationships between the different indices. The raw score correlation between judge scores and empirical p-values in the first column serves as a useful 'standard' for judging how well indices measure accuracy. In the internal framework, all of the correlational indices are in good agreement with the score/p-value correlations. The residual-based indices again show a weaker relationship with the score/p-value correlations, although the relationship is stronger than it was for reading. This changes sharply in the external framework. Since the score-expected and point-measure correlations are now based on the same data as the score/p-value correlationss, it is not surprising that they show a near-perfect correlation. More interesting is that the fit indices now correlate very closely with the score/p-value correlations - nearly as closely as the correlational indices do. Overall, the pattern is the same as for listening: the fit indices are much more sensititive to inaccuracy in the external framework.

Table 4.9. *Correlation Matrices of Inaccuracy Indices for Listening*

	Corr P-values	SR/ROR	Score-Expected	Point-Measure	SD Residuals	InFit MnSq	InFit Z	OutFit MnSq
INTERNAL								
SR/ROR	.747**							
Score-Expected	.766**	.998**						
Point-Measure	-.761**	-.999**	-.999**					
SD Residuals	-.480*	-.816**	-.788**	-.797**				
InFit MnSq	-.557*	-.843**	-.819**	.827**	.983**			
InFit Z	-0.468	-.794**	-.766**	.776**	.985**	.987**		
OutFit MnSq	-.566*	-.860**	-.837**	.845**	.984**	.999**	.985**	
OutFit Z	-.472*	-.808**	-.780**	.790**	.988**	.985**	.998**	.986**
EXTERNAL								
SR/ROR	.747**							
Score-Expected	.986**	.715**						
Point-Measure	.983**	.728**	.998**					
SD Residuals	-.919**	-.779**	-.920**	-.925**				
InFit MnSq	-.955**	-.799**	-.958**	-.960**	.981**			
InFit Z	-.956**	-.729**	-.960**	-.962**	.981**	.988**		
OutFit MnSq	-.959**	-.823**	-.961**	-.967**	.983**	.995**	.983**	
OutFit Z	-.960**	-.756**	-.961**	-.966**	.983**	.982**	.994**	.988**

* Significant at 0.05 level (2-tailed). **Significant at 0.01 level (2-tailed).

Figure 4.8 allows us to visually examine these relationships. For the score-expected correlations, we see the same reasonably accurate linear relationship but generally higher level of correlations in the internal framework.

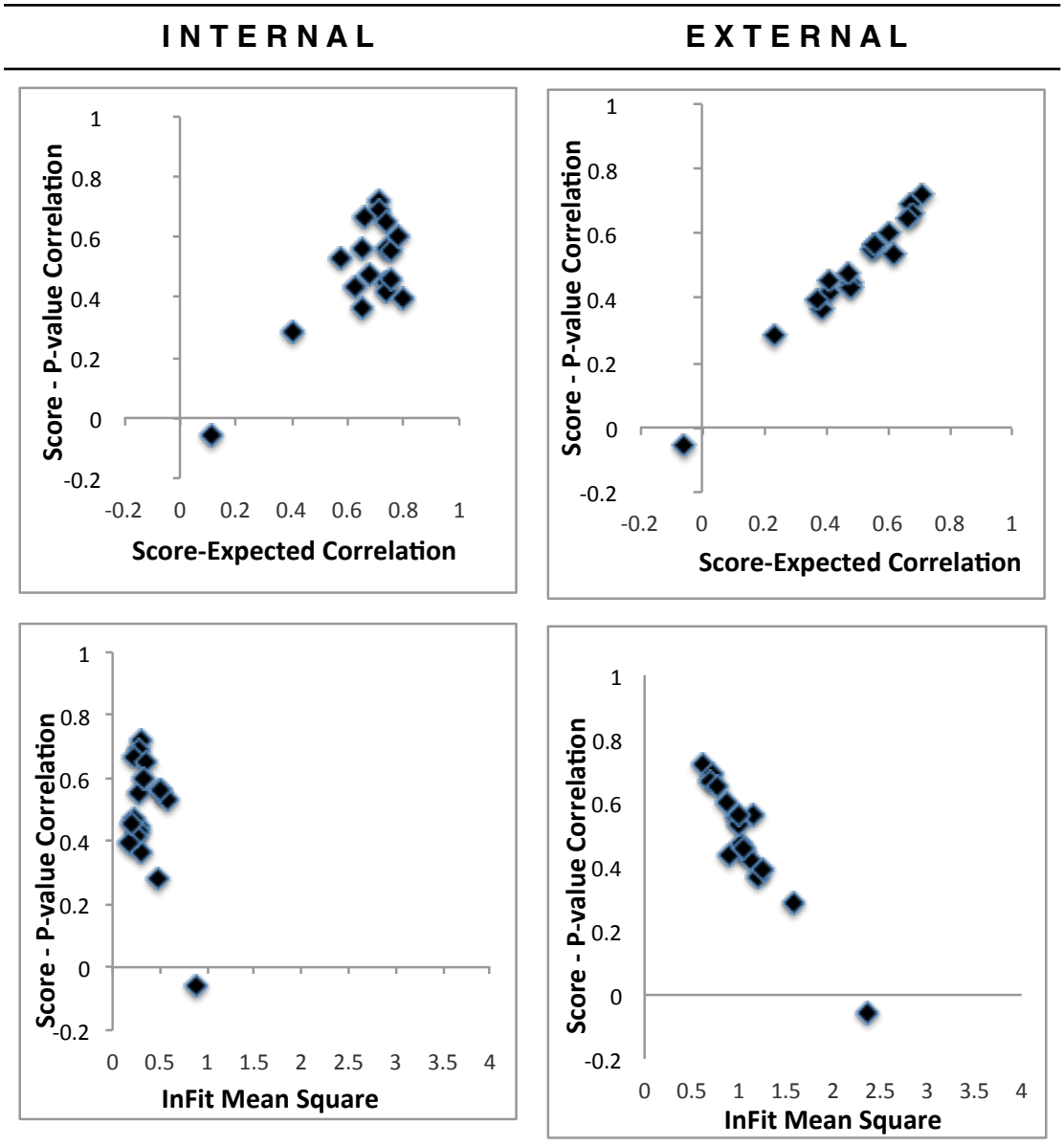


Figure 4.8. Indices v. score/p-value correlations, listening, internal v. external.

For the residual-based indices, values again appear relatively compressed and this, in turn, seems to suppress correlations with judge accuracy, which appear very clearly in the external framework.

Group-level Effects

Item separation statistics have been proposed for examining group-level inaccuracy effects, with a lack of separability suggesting inaccuracy. The key findings are summarized here.

Reading. The item separation statistics in Table 4.10 suggest a low level of separation (fixed chi sq = 271.1, separation ratio = 2.44, strata = 3.59, and reliability of the separation index was 0.86), consistent with group-level inaccuracy. While the chi-square test did show statistically significant separation, this is unsurprising given a forty-item test. The existence of fewer than four distinct strata or levels of item difficulty suggests a lack of discrimination. Comparison with the external frame reveals a strong contrast. The indicators from the external frame differ markedly: the fixed chi square value (1129.5, v. 271.1), the separation ratio (5.40, v. 2.44), the number of statistically distinct strata (7.53, v. 3.59) and the reliability of the separation index (0.97, v. 0.86) all suggest considerably more dispersion of items in the external framework. This is consistent with a group-level inaccuracy effect making it more difficult to discriminate between different items. However, group-level centrality might also explain this result. This will be returned to below.

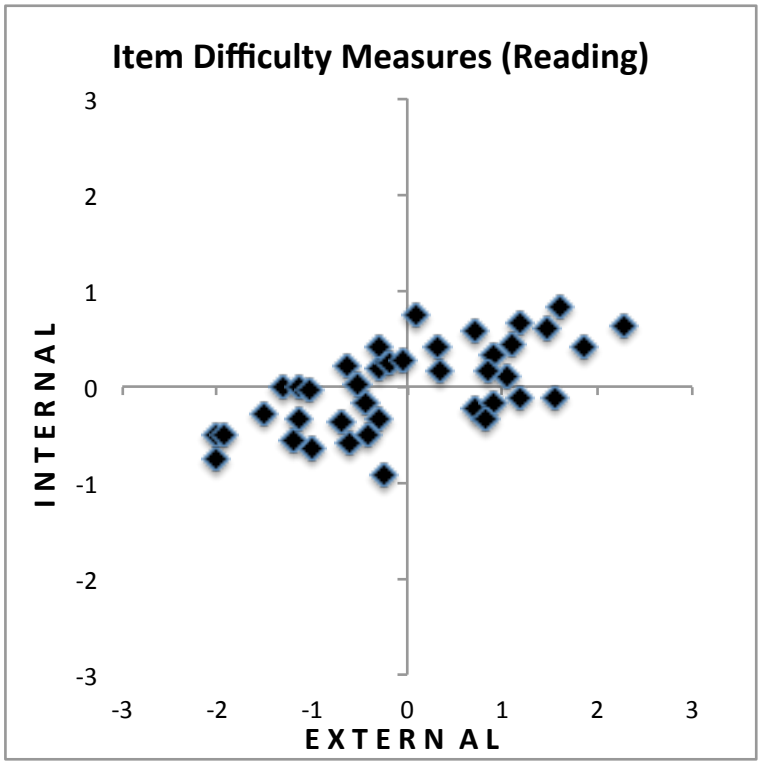
Table 4.10. *Item Separation Statistics, Internal v. External*

	INTERNAL	**EXTERNAL**
READING		
Fixed Chi Square	271.1, df = 39, p = .00	1129.5, df = 39, p = .00
Separation Ratio	2.44	5.40
Separation Index (Strata)	3.59	7.53
Reliability of the Separation Index	0.86	0.97
LISTENING		
Fixed Chi Square	180.5, df = 39, p = .00	794.8, df = 39, p = .00
Separation Ratio	1.89	4.45
Separation Index (Strata)	2.85	6.27
Reliability of the Separation Index	0.78	0.95

Listening. The item separation statistics for listening in Table 4.10 show an even smaller degree of separation than did the reading results and are also consistent with a group-level inaccuracy effect. Comparing results across the two frames, all indices - the fixed chi square value (794.8, v. 180.5), the separation ratio (4.45, v. 1.89), the number of statistically distinct strata (6.27, v. 2.85) and the reliability of the separation index (0.95, v. 0.78) - show less separation in the internal frame, consistent with group-level inaccuracy.

Thus, for both reading and listening, we find a relative lack of separation in the internal frame. However, directly comparing the item measures in Figure 4.9, it is not clear that group-level inaccuracy is responsible for this lack of separation in the internal frame. The linear relationship between the variables seems reasonably high (for reading, $r = 0.64$; for listening, $r = 0.75$), but it is very clear that the range of measures in the internal frame is severely compressed in relation to those in the external frame. This question will be returned to in the next section, on centrality/ extremity.

Summary. Results for the internal frame did not match those for the external frame, with results for the internal frame appearing overly lenient or 'optimistic' and failing to flag a number of judges who were flagged within the external frame. The *group-level indicators within the internal frame* were consistent with a group-level inaccuracy effect.



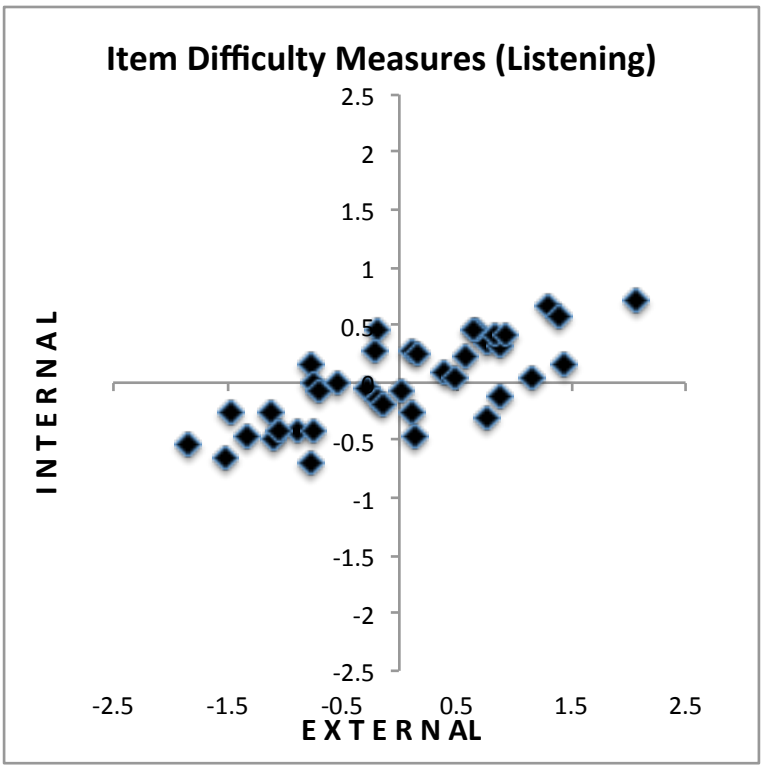


Figure 4.9. Item difficulty in logits for reading and listening, internal v. external.

4.1.3 Centrality/Extremity

Individual-level Effects

The residual-expected and residual-measure correlations yielded essentially identical results, as did the residual-based indices (standard deviation of the residuals, fit statistics). Thus, to simplify the presentation of results, only results for the residual-expected correlations, infit mean square statistics and the raw score standard deviations are shown below. Results for all indicators can be found in Appendix D. Values for all indices also appear in the correlation matrices below, which indicate the high level of agreement found.

Reading. Table 4.11 displays the standard deviation of 'raw score' estimates and the MFRM indices of centrality for each judge. In the internal frame, the residual-expected correlations flag judges J02, J14, J15, J16 and J18 for displaying extremity effects and judges J05, J08, J10 and J12 for centrality. The infit mean square values indicate overfit to the model for J01, J02, J05, J08, J09, J10 and J13. No judges were flagged for misfit.

Results differ markedly in the external framework. All judges have negative correlations, suggesting centrality, and the correlations are significant at the .05 level for all judges except for J03 and J18. Looking at the infit mean square values, all judges have values above 1.0, and only J13 is not flagged for misfit or extremity.

Table 4.11. *Indices of Centrality/Extremity for Reading, Internal v. External*

		$r_{exp,res}$		InFit Mean Square	
Judge	Raw Score Standard Deviation	INTERNAL	EXTERNAL	INTERNAL	EXTERNAL
J01	11.9	0.08	-0.62*	0.56*	1.87**
J02	15.4	0.36**	-0.55*	0.39*	1.55**
J03	17.8	0.24	-0.30	1.11	1.89**
J04	11.2	-0.31	-0.72*	0.63	1.77**
J05	9.6	-0.51*	-0.84*	0.42*	1.64**
J06	13.7	-0.19	-0.70*	0.70	2.36**
J07	14.7	-0.31	-0.76*	1.01	3.58**
J08	9.8	-0.39*	-0.77*	0.56*	1.73**
J09	12.2	-0.14	-0.71*	0.43*	1.66**
J10	8.3	-0.62*	-0.88*	0.55*	2.33**
J11	13.2	-0.28	-0.77*	0.69	2.78**
J12	11.9	-0.32*	-0.75*	0.61	2.03**
J13	10.7	-0.20	-0.73*	0.35*	1.05
J14	16.3	0.39**	-0.45*	0.57*	1.64**
J15	19.4	0.51**	-0.43*	0.81	2.65**
J16	19.7	0.55**	-0.33*	0.81	2.07**
J17	16.3	0.27	-0.58*	0.57*	2.28**
J18	20.1	0.59**	-0.23	0.78	1.55**
*Overfit/ Centrality		4	16	9	0
No Effect		9	2	9	1
**Underfit/ Extremity		5	0	0	17

*Negative correlation, $p < .05$; or, infit mn. sq. < .6.
**Positive correlation, $p < .05$; or, infit mn. sq. > 1.4.

Figure 4.10 provides a visual comparison of results across both frameworks. The values of the correlational indices are shifted to all negative correlations in the external framework. As was observed above, the fit values appear relatively 'compressed' in the internal framework.

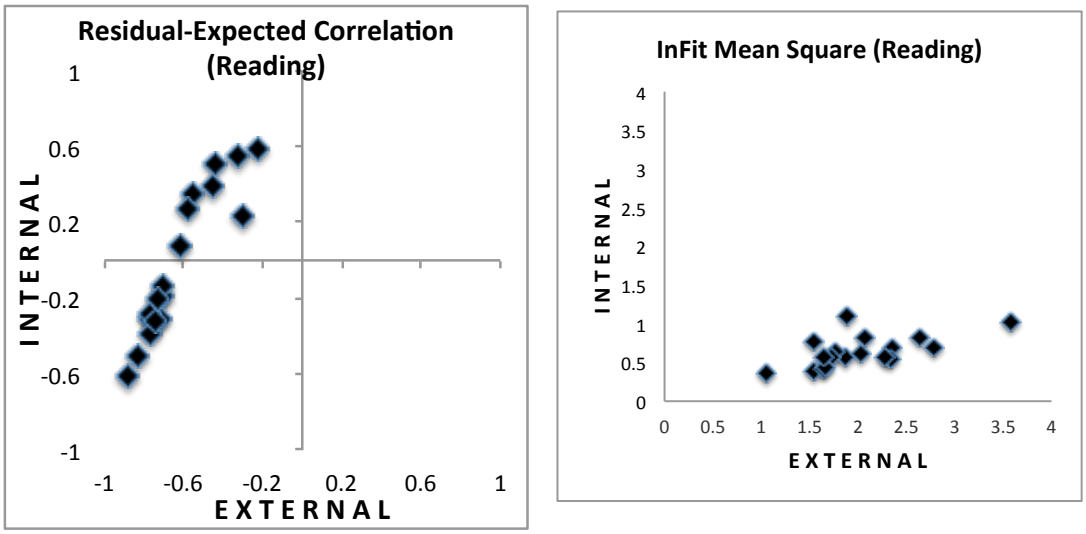


Figure 4.10. Centrality/extremity indices for reading, internal v. external.

Using the correlation matrices in Table 4.12 to compare the correlational with the residuals-based indices, note that while there is some overlap in the internal framework, they move in opposite directions within the external framework. Using the raw score standard deviation, in the first column, as an indicator of centrality in the external framework, we can see that the correlational indices are in good agreement with it in both frameworks. In the internal framework, there are reasonably strong positive correlations between the raw-score standard deviation and the residuals-based indices, meaning that low values for both was suggestive of centrality. However, in the external framework, this relationship nearly disappears: the correlations are low and non-significant.

Table 4.12. Correlation Matrices of Centrality/Extremity Indices for Reading

	Raw Score SD	SD Residuals	InFit MnSq	InFit Z	OutFit MnSq	OutFit Z	Res-Exp Corr
INTERNAL							
SD Residuals	.624**						
InFit MnSq	.587*	.966**					
InFit Z	.578*	.968**	.992**				
OutFit MnSq	.588*	.985**	.992**	.988**			
OutFit Z	.578*	.983**	.984**	.994**	.993**		
Res-Exp Corr	.915**	0.289	0.290	0.283	0.270	0.264	
Res-Meas Corr	.913**	0.287	0.288	0.282	0.269	0.263	1.000**
EXTERNAL							
SD Residuals	0.220						
InFit MnSq	0.161	.981**					
InFit Z	0.179	.983**	.992**				
OutFit MnSq	0.029	.973**	.981**	.971**			
OutFit Z	0.035	.976**	.975**	.980**	.992**		
Res-Exp Corr	.911**	-0.138	-0.176	-0.150	-0.318	-0.306	
Res-Meas Corr	.901**	-0.148	-0.176	-0.150	-0.325	-0.311	.997**

* Significant at 0.05 level (2-tailed). **Significant at 0.01 level (2-tailed).

The scatterplots in Figure 4.11 suggest that in the external framework, when the residuals are less 'compressed', the correlation with the raw score standard deviation disappears (recall from above, that in the external framework, the residuals-based indices appeared to become more sensitive to *inaccuracy* and correlated more strongly with the *score/p-value correlations*).

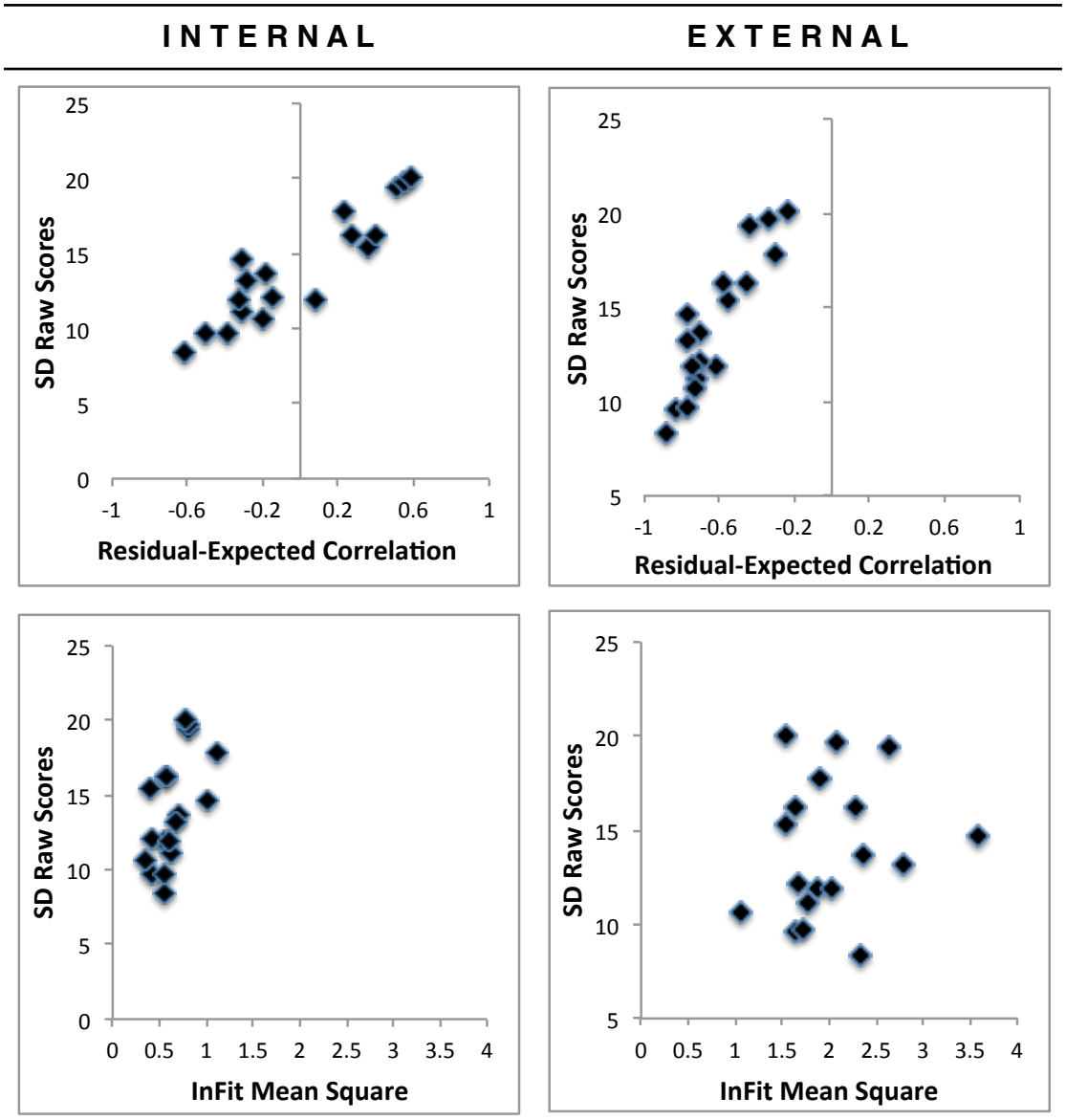


Figure 4.11. Centrality/extremity indices v. raw score standard deviations, reading, internal v. external.

Listening. Table 4.13 displays the standard deviation of 'raw score' estimates and the MFRM correlational indices of centrality for each judge. In the internal frame, the expected-residual correlation flags judges J04 and J16 for extremity and J07 and J11 for centrality. For the infit mean square values in Table 4.13, 17 judges overfit the model, with J11 being the only judge showing acceptable fit. In the external framework, results again sharply differ. The expected-residual correlation flags *all* judges for centrality. The infit mean square values, however, flag only two judges (J07 and J11) for misfit or underfit. The other 16 judges had acceptable fit values.

Table 4.13. *Indices of Centrality/Extremity for Listening, Internal v. External*

		$r_{exp,res}$		InFit Mean Square	
Judge	Raw Score Standard Deviation	INTERNAL	EXTERNAL	INTERNAL	EXTERNAL
J01	11.0	0.15	-0.64*	0.29*	1.06
J02	10.0	0.09	-0.61*	0.30*	0.61
J03	8.1	-0.08	-0.70*	0.28*	0.89
J04	14.1	0.34**	-0.45*	0.51*	1.14
J05	8.9	-0.05	-0.72*	0.31*	1.20
J06	8.6	-0.19	-0.78*	0.22*	1.02
J07	10.0	-0.37*	-0.79*	0.47*	1.57**
J08	8.7	0.11	-0.69*	0.25*	1.12
J09	11.1	0.13	-0.65*	0.28*	0.98
J10	13.7	-0.03	-0.55*	0.57*	0.99
J11	11.4	-0.45*	-0.78*	0.89	2.37**
J12	10.2	0.05	-0.65*	0.28*	0.71
J13	8.6	-0.19	-0.76*	0.23*	0.69
J14	12.2	0.15	-0.50*	0.50*	1.01
J15	12.2	0.22	-0.53*	0.36*	0.76
J16	12.1	0.32**	-0.52*	0.34*	0.87
J17	9.9	0.13	-0.74*	0.19*	1.25
J18	9.4	0.06	-0.73*	0.20*	1.06
*Overfit/ Centrality		2	18	17	0
No Effect		14	0	1	16
*Underfit/ Extremity		2	0	0	2

*Negative correlation, $p < .05$; or, infit mn. sq. < .6.
**Positive correlation, $p < .05$; or, infit mn. sq. > 1.4.

Figure 4.12 provides a visual comparison of results across both frameworks. As with reading, the values of the correlational indices are shifted to all negative correlations in the external framework. As was observed above, the fit values were again relatively 'compressed' in the internal framework.

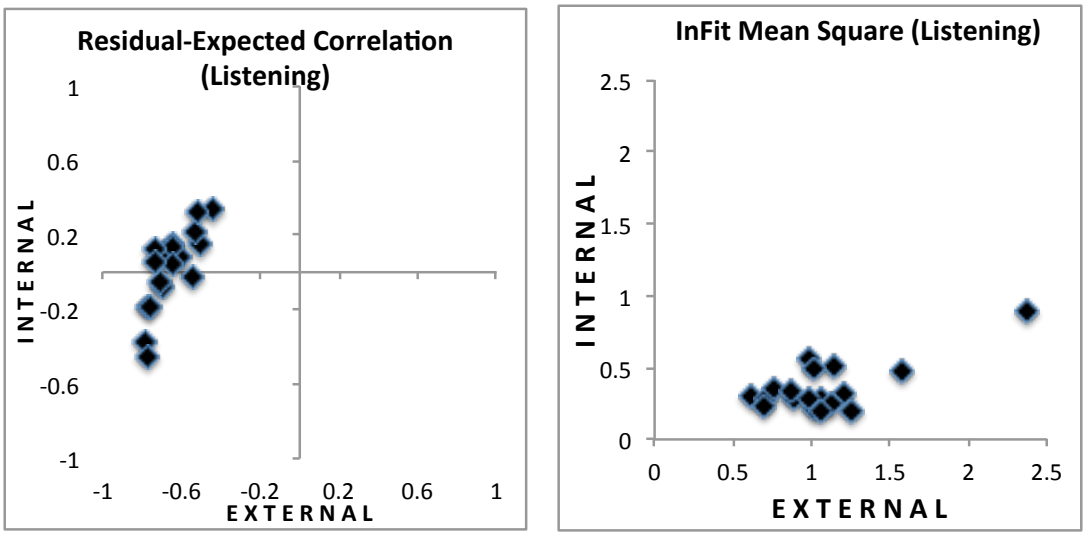


Figure 4.12. Centrality/extremity indices for listening, internal v. external.

The correlational matrices in Table 4.14 allows closer investigation of how the different indices performed in the two frameworks. Comparing the correlational with the residuals-based indices, the residuals-based indices actually correlate more strongly with the raw score standard deviations than do the correlations between residuals and expected scores and measures. The correlational indices have positive non-significant correlations with the standard deviations. However, in the external framework, this situation is reversed. The correlations between SD Residuals and the raw score standard deviations, and the fit values and the raw score standard deviations falls below significance to near zero, while the residual-expected and residual measure correlations approach 0.8, and are significant at the .01 level.

Table 4.14. *Correlation Matrices of Centrality/Extremity Indices for Listening*

	Raw Score SD	SD Residuals	InFit MnSq	InFit Z	OutFit MnSq	OutFit Z	Res-Exp Corr
INTERNAL							
SD Residuals	.639**						
InFit MnSq	.585*	.983**					
InFit Z	.636**	0.985	.987**				
OutFit MnSq	.559*	0.984	.999**	.985**			
OutFit Z	.616**	0.988	.985**	.998**	.986**		
Res-Exp Corr	0.425	-0.360	-0.371	-0.300	-0.405	-0.329	
Res-Meas Corr	0.417	-0.368	-0.378	-0.308	-0.411	0.337	-.999**
EXTERNAL							
SD Residuals	0.155						
InFit MnSq	0.092	.981**					
InFit Z	0.092	.981**	.988**				
OutFit MnSq	0.055	.983**	.995**	.983**			
OutFit Z	0.053	.983**	.982**	.994**	.988**		
Res-Exp Corr	.796**	-0.379	-0.395	-0.381	-0.441	-0.432	
Res-Meas Corr	.781**	-0.372	-0.379	-0.364	0.432	-0.422	.995**

* Significant at 0.05 level (2-tailed). **Significant at 0.01 level (2-tailed).

Figure 4.13 depicts the same general pattern noted above. Residuals are compressed within the internal framework; in the external framework, the correlation with the raw score standard deviations tends to disappear for the fit values while strengthening for the residual expected correlations. Examining the plot, the lower-than-expected correlations between the residual-expected and the raw score standard deviations may have resulted from a small number of outlying observations.

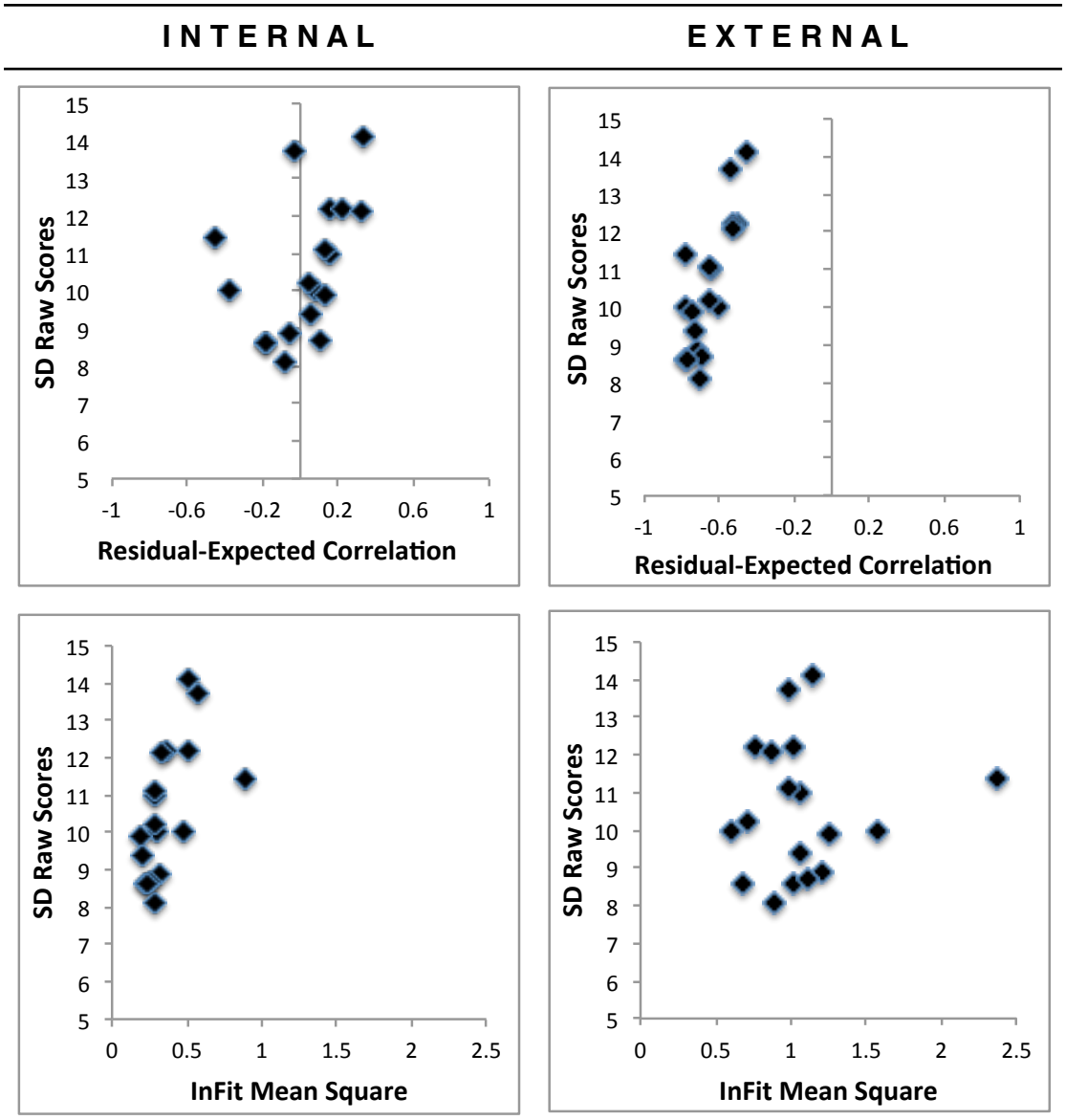


Figure 4.13. Centrality/extremity indices v. raw score standard deviations, listening, internal v. external.

Group-level Effects

Item separation statistics and item fit indices have been suggested for detecting group-level centrality effects. As item separation statistics are also indicators of group-level inaccuracy, they were presented above. There it was noted that the items were clearly less separated in the internal framework and that this was consistent with a group-level inaccuracy effect. However, this outcome is also consistent with a group-level centrality effect. Given the outcomes for individual judges on the centrality indices above, it seems likely that, in fact, the low level of separation was due to a group-level centrality effect.

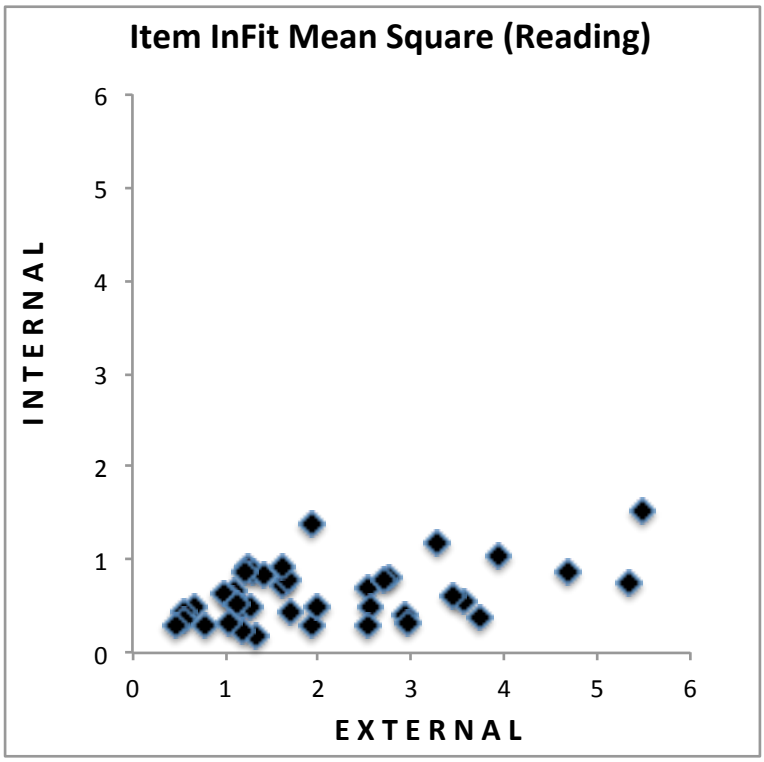
It was also suggested that item mean square values significantly below 1.0 might indicate a group-level centrality effect. Tables 4.15 and 4.16 compare the item infit mean square values across frameworks for listening and reading. It can be seen that a large number of items have infit values below 1.0 (36 of 40 for reading, and 40 of 40 for listening). From Figure 4.14, the same pattern of highly compressed values in the internal framework is once again seen. These results are consistent with a group-level centrality effect.

Table 4.15. *Item Fit Indices for Reading, Internal v. External*

	INTERNAL		EXTERNAL			INTERNAL		EXTERNAL	
Items	Measure	InFit MS	Measure	InFit MS	Items	Measure	InFit MS	Measure	InFit MS
iR01	0.11	0.45	1.04	1.69	iR21	-0.22	0.49	0.72	1.98
iR02	0.42	0.69	-0.31	2.52	iR22	0.60	0.49	1.48	1.26
iR03	-0.28	0.42	-1.52	2.93	iR23	0.75	1.17	0.10	3.28
iR04	0.00	0.38	-1.32	3.73	iR24	0.63	1.53	2.28	5.49
iR05	0.16	0.93	0.35	1.24	iR25	0.42	1.05	1.85	3.93
iR06	-0.58	0.48	-0.61	0.66	iR26	-0.52	0.87	-2.00	4.68
iR07	0.22	0.82	-0.63	2.76	iR27	-0.34	0.43	-0.29	0.56
iR08	0.65	0.88	1.20	1.29	iR28	-0.52	0.54	-1.92	3.55
iR09	0.42	0.68	0.32	1.10	iR29	-0.65	0.36	-0.99	0.54
iR10	0.58	1.37	0.72	1.93	iR30	0.00	0.60	-1.15	3.46
iR11	-0.75	0.29	-2.01	1.92	iR31	-0.92	0.19	-0.24	1.34
iR12	0.19	0.71	-0.31	1.62	iR32	-0.37	0.28	-0.70	0.48
iR13	0.32	0.57	0.91	1.08	iR33	-0.55	0.53	-1.19	1.13
iR14	0.45	0.83	1.11	1.42	iR34	-0.16	0.77	0.90	2.71
iR15	-0.16	0.63	-0.43	0.98	iR35	-0.52	0.86	-0.40	1.20
iR16	0.24	0.77	-0.19	1.66	iR36	-0.03	0.50	-1.03	2.57
iR17	-0.11	0.32	1.18	2.96	iR37	0.27	0.30	-0.06	0.79
iR18	0.16	0.48	0.86	1.19	iR38	-0.34	0.25	-1.15	1.19
iR19	0.83	0.93	1.62	1.62	iR39	-0.34	0.30	0.81	2.52
iR20	-0.11	0.75	1.55	5.35	iR40	0.03	0.31	-0.51	1.03

Table 4.16. *Item Fit Indices for Listening, Internal v. External Frames*

	INTERNAL		EXTERNAL			INTERNAL		EXTERNAL	
Items	Measure	InFit MS	Measure	InFit MS	Items	Measure	InFit MS	Measure	InFit MS
iL1	0.08	0.28	0.38	0.41	iL21	-0.65	0.28	-1.52	1.21
iL2	0.16	0.25	-0.77	2.11	iL22	-0.43	0.3	-0.88	0.65
iL3	-0.01	0.39	-0.75	1.53	iL23	-0.5	0.28	-1.11	0.81
iL4	0.34	0.31	0.76	0.52	iL24	-0.04	0.33	-0.28	0.56
iL5	0.47	0.42	-0.2	1.78	iL25	0.22	0.59	0.57	0.81
iL6	0.28	0.36	0.11	0.61	iL26	-0.26	0.26	0.11	0.48
iL7	-0.26	0.43	-1.47	2.84	iL27	0.58	0.42	1.4	1.14
iL8	-0.07	0.19	-0.71	0.93	iL28	0.42	0.24	0.92	0.48
iL9	-0.69	0.07	-0.78	0.09	iL29	-0.43	0.21	-0.75	0.37
iL10	-0.26	0.3	-1.12	1.41	iL30	-0.3	0.36	0.76	2
iL11	0.31	0.28	0.89	0.65	iL31	0.05	0.6	1.15	2.34
iL12	-0.01	0.4	-0.55	1.08	iL32	-0.47	0.13	-1.33	1.03
iL13	0.25	0.44	0.16	0.64	iL33	0.45	0.34	0.65	0.41
iL14	0.72	0.66	2.06	2.93	iL34	-0.14	0.42	-0.2	0.53
iL15	0.28	0.59	-0.21	1.41	iL35	0.16	0.48	1.43	2.71
iL16	0.66	0.88	1.3	1.39	iL36	0.05	0.56	0.48	0.83
iL17	-0.07	0.44	0.02	0.53	iL37	-0.54	0.08	-1.84	2.09
iL18	0.47	0.32	0.67	0.4	iL38	-0.43	0.25	-1.06	0.81
iL19	0.42	0.3	0.83	0.47	iL39	-0.47	0.18	0.14	0.76
iL20	-0.11	0.29	0.88	1.68	iL40	-0.2	0.26	-0.14	0.34



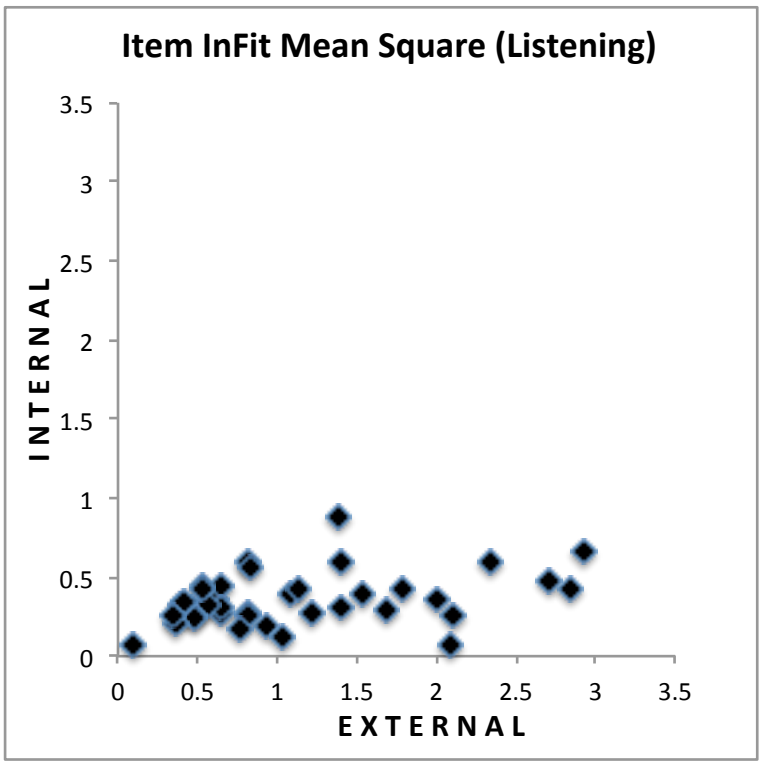


Figure 4.14. Item infit mean square values in logits for reading and listening, internal v. external frameworks.

Summary. The results for the internal and external frames clearly do not match, with the internal frame failing to flag a large number of judges showing a centrality bias,

for both the listening and the reading exams. Indeed, judges flagged for possibly demonstrating an 'extremity' bias appeared to display centrality within the external frame. Importantly, however, the *group-level indicators within the internal frame* did appear to correctly indicate the presence of a group-level centrality effect.

4.1.4 Summary

Above, it was seen that results from the internal and external frameworks did not agree. However, within the frameworks, different indices gave different results. Before summarizing the results for the two frameworks, it is first necessary to decide which indicators to use, based on correlations with the raw score statistics and findings from earlier research. For the *external* framework, it seems reasonable to use the score-expected and point-measure correlations as the MFRM indicators for inaccuracy and the residual-expected and residual-measure correlations as the MFRM indicators for centrality/extremity. For borderline cases, where these indicators reached different decisions, the raw score statistic (score/p-value correlation for inaccuracy and raw score standard deviation for centrality/extremity) was given the deciding 'vote.' For the internal framework, the SR/ROR correlation was in closest agreement with the raw score indicator and was used to indicate inaccuracy. For centrality/extremity, the residual-expected and measure-residual correlations were used for centrality/extremism for the reading test, and the SD of the residuals was used for the listening test. (Again, note that some of these indicators were not shown above; Appendix D provides the values for all indicators.)

The final results are summarized in Table 4.17. There are clear differences between the results for the internal and external frameworks, suggesting that the assumption of the MFRM for detecting rater effects in a standard setting situation using internal data were not met, and that the model was not robust to violations of this assumption. On a more positive note, the item separation and fit statistics were sensitive to the presence of group-level effects.

Table 4.17. *Summary of Flagged Raters, Reading and Listening, Internal v. External*

	READING		**LISTENING**	
	Internal	**External**	**Internal**	**External**
Leniency	J11, J07	J11, J07	J07, J10	J07, J10
Severity	J01, J08	J01, J08		J03
Inaccuracy	J04, J07, J08, J10, J11, J12	J01, J04, J06, J07, J08, J10, J11, J12, J17	J07, J11	J05, J07, J11, J17
Centrality	J05, J08, J10, J12	J01, J02, J04, J05, J06, J07, J08, J09, J10, J11, J12, J13, J14, J15, J16, J17, J18	J01, J02, J03, J06, J08, J09, J12, J13, J17, J18	J01, J02, J03, J04, J05, J06, J07, J08, J09, J10, J11, J12, J13, J14, J15, J16, J17, J18
Extremism	J02, J14, J15, J16, J18		J11	

4.2 Assumptions of the Angoff Method

Table 4.18 provides raw score summaries of judge performance, along with indicators showing which judges were flagged for rater effects in the above analysis.

Table 4.18. *Raw Score Statistics and Summary of Flagged Raters, Reading and Listening*

		READING			LISTENING		
Group	Judge	Mean*	SD**	*r****	Mean*	SD**	*r****
I (Mon)	J01	85.3 - S	11.9 - C	0.33 - I	79.0	11.0 - C	0.45
	J02	71.3	15.4 - C	0.54	78.6	10.0 - C	0.72
	J03	77	17.8	0.57	85.2 - S	8.1 - C	0.43
	J04	76.8	11.2 - C	0.37 - I	77.1	14.1 - C	0.56
	J05	69.6	9.6 - C	0.43	79.3	8.9 - C	0.37 - I
	J06	72.4	13.7 - C	0.21 - I	76.5	8.6 - C	0.47
II (Wed)	J07	53.9 - L	14.7 - C	-0.04 - I	66.9 - L	10.0 - C	0.28 - I
	J08	81.4 - S	9.8 - C	0.31 - I	81.5	8.7 - C	0.42
	J09	71.0	12.2 - C	0.43	71.8	11.1 - C	0.55
	J10	70.4	8.3 - C	0.06 - I	53.6 - L	13.7 - C	0.53
	J11	44.5 - L	13.2 - C	0.12 - I	74.2	11.4 - C	-0.06 - I
	J12	71.6	11.9 - C	0.30 - I	73.9	10.2 - C	0.69
III (Fri)	J13	75.6	10.7 - C	0.63	76.1	8.6 - C	0.66
	J14	73.5	16.3 - C	0.54	79.5	12.2 - C	0.56
	J15	63.5	19.4 - C	0.41	74.9	12.2 - C	0.65
	J16	69.3	19.7 - C	0.55	79.8	12.1 - C	0.60
	J17	62.8	16.3 - C	0.38 - I	76.6	9.9 - C	0.39 - I
	J18	64.8	20.1 - C	0.65	83.9	9.4 - C	0.46
Judge Mean		69.7	14.0	0.38	76.0	10.6	0.49
p-values		65.4	21.7	1.00	67.3	18.7	1.00

*Mean probability estimate. In this column, 'L' = Lenient and 'S' = Severe.
**Raw score standard deviations. In this column, 'C' = Centrality and 'E' = Extremity.
***Correlation of estimates with p-values. In this column, 'I' = Inaccurate.

Assumption of Accurate Representation of the Barely Proficient Student

This was evaluated by evaluating the hypothesis that leniency/severity effects existed which would have violated the assumption. In terms of correctly identifying the B1 cut score level, the most that can be said is that only a small number of judges seemed either too lenient or too severe and that the remaining judges set the cut score at a level that does not seem unreasonable.

Assumption of Accurate Representation of Item Functioning

In terms of accuracy, it is clear that a number of judges had difficulty in making their judgments based solely on the latent trait (or 'construct') in question. The group as a whole, however, outperformed the individual judges. The correlation of the group mean estimates for each item and the empirical p-values was 0.62 for reading and 0.74 for reading within the raw score framework (corresponding values for item measures within the Rasch framework were 0.64 for reading and 0.75 for listening). Performance thus improved in the listening study, which was held after all three rounds of the reading study had been completed, suggesting a practice effect.

Quantification Assumption

The third and final assumption is that the Angoff judges can quantify their predictions using the 0 to 1 probability scale. The above finding of a group-level centrality effect shows that this assumption was not met. For the reading test, 17 of 18 judges displayed a centrality effect, and all 18 did so for the listening test. This resulted in a scale that was considerably more compressed than was the original scale. The standard deviation serves as a good indicator of the extent of this scale compression. The standard deviations for all items on the reading and listening forms were 21.7 and 18.7 respectively. Using conditional probabilities based on the B1 cut score from the first round, the corresponding standard deviations would be 21.3 and 16.2. Based on the estimates of the Angoff judges, the standard deviations were 8.9 and 7.0 for reading and listening, respectively. Clearly, this indicates a severe compression of the metric.

Results Across Three Rounds

The above analysis focused on only the first round of the standard setting. Table 4.19 allows us to examine how raw score indicators of inaccuracy and centrality/extremity changed across all three rounds. (The values based on conditional probabilities in Table 4.19 are the expected values for students *at the cut score selected by the group of judges as a whole*; the slight difference in the mean percentage correct for the conditional probabilities versus the judge estimates is a result of translation from the Rasch (external frame) to the raw score metric when the conditional probabilities

were calculated). Recalling that judges were presented with empirical p-values between rounds, it is clear that this seems to have improved their accuracy: the correlation between the mean of judges' item estimates and empirical p-values reached 0.92 for reading and 0.96 for listening by the third round. However, exposure to the empirical data was less beneficial in terms of the centrality effect: even by the end of the third round, the standard deviation of the judges' estimates was only 57% of its expected size based on the conditional probabilities for reading, and 72% of its expected size for listening. This indicates that the judge-constructred scale remained compressed even following exposure to empirical data.

Table 4.19. *Raw Score Statistics For All Rounds, Reading and Listening*

	Round 1			**Round 2**			**Round 3**		
	Mean	SD	Corr with Item p-values	Mean	SD	Corr with Item p-values	Mean	SD	Corr with Item p-values
READING									
Item *p*-values	65.4	21.7	1.00	65.4	21.7	1.00	65.4	21.7	1.00
Conditional Probabilities	70.3	20.3	1.00	74.2	18.8	1.00	74.9	18.5	1.00
Judge Estimates	69.7	8.9	0.62	74.0	10.2	0.86	74.6	10.6	0.92
LISTENING									
Item *p*-values	67.3	18.7	1.00	67.3	18.7	1.00	67.3	18.7	1.00
Conditional Probabilities	76.5	15.3	0.99	76.3	15.3	0.99	77.6	14.8	0.99
Judge Estimates	76.0	7.0	0.74	76.0	9.7	0.93	77.2	10.6	0.96

Impact on Cut Scores and Pass/Fail Rates

Estimates of the impact of these distortions of the scale in the internal frame on cut scores and pass/fail rates can be seen in Table 4.20. The compressed scale 'pulled' the cut score closer to the mean for both tests, thus lowering the cut score and increasing the pass rate. Even by the end of the third round, an additional 10.5% of all students were counted as reaching the B1 cut score for the reading test, with an additional

4.4% passing the listening test.

Table 4.20. *Cut Scores*, Standard Deviations and Pass Rates** for All Rounds*

	Round 1			Round 2			Round 3		
	Cut Score (logits)	SD (logits)	Pass Rate	Cut Score (logits)	SD (logits)	Pass Rate	Cut Score (logits)	SD (logits)	Pass Rate
READING									
Internal	0.91	0.46	46.4%	1.13	0.56	37.8%	1.17	0.58	37.7%
External	1.09	1.11	38.1%	1.33	1.11	29.0%	1.37	1.11	27.2%
LISTENING									
Internal	1.23	0.40	35.4%	1.25	0.56	35.4%	1.34	0.61	31.5%
External	1.39	0.93	31.0%	1.38	0.93	31.0%	1.46	0.93	27.1%

*Estimated by Winsteps using a 100-point binomial traits model; results may differ slightly from the 11-point binomial model estimated with Facets for the earlier analyses of rater effects.
**Based on the samples used in the original calibration of the exams (n = 13,131 for reading; n = 13,012 for listening).

Rater Background Characteristics

Finally, in terms of rater background characteristics that might influence the likelihood of demonstrating rater effects, Table 4.21 shows the average cut scores (severity measures, in logits), mean score-expected correlations (accuracy) and residual-expected correlations for English native speakers versus non-native speakers, and judges with an administrative role versus judges with a teaching role. Non-native speakers and adminstrators were more severe, more accurate and less prone to display a centrality bias. However, with the very small number of judges involved (four NS judges; 3 judges with administrative roles), little can be made of this observation.

Table 4.21. *Judge Characteristics and Indices for Severity, Accuracy and Centrality*

	READING			LISTENING		
	Mean B1 Cut Score (Judge Severity Measure)	**Mean Score-Expected Corr.**	**Mean Res-Expected Corr.**	**Mean B1 Cut Score (Judge Severity Measure)**	**Mean Score-Expected Corr.**	**Mean Res-Expected Corr.**
English Native Speakers*	0.46	0.19	-0.78	1.25	0.24	-0.74
English Non-native Speakers	1.11	0.43	-0.57	1.33	0.56	-0.63
Teachers	0.91	0.37	-0.65	1.28	0.47	-0.67
Admin. Role**	1.23	0.43	-0.46	1.47	0.58	-0.57

*NS (n = 4): J04, J07, J08, J12 **Admin. Role (n = 3): J01, J13, J14.

CHAPTER FIVE DISCUSSION AND CONCLUSION

The two-fold purpose of this study was, first, to evaluate the assumption underlying use of latent trait models for detecting rater effects, and, second, to evaluate whether the assumptions underlying the use of the Angoff standard setting method held in a particular case. The main finding is that the assumptions were found *not* to hold. The most striking departure from these assumptions was related to a group-level *centrality* or *central tendency* effect. This chapter summarizes and discusses the theoretical and practical importance of the results presented above. Suggestions for practitioners are offered, limitations of the study are noted, and future research directions are suggested.

5.1 Summary of Results

The study used data from an operational standard setting study conducted to set cut scores linking English as a Foreign Language reading and listening exams of forty items each to the Common European Framework of Reference. The data consisted of the empirical item-response data from the original administration of the exams to students at a private university in Taiwan, and of the estimates made in the first round of a modified Angoff standard setting by 18 trained EFL professionals of the probability that a barely proficient CEFR B1 student would answer each item correctly.

Assumption of Latent Trait Models for Detecting Rater Effects

Latent trait models generally, and the many-facet Rasch model in particular have come into widespread use for the detection of rater biases or effects. Generally, the use of internal data alone (i.e., from the ratings of the judges) is used for this purpose. A number of researchers have proposed using the MFRM for evaluating standard setting results and the ability of the model to detect biases based solely on internal data has been noted as an advantage (Engelhard & Anderson, 1998). However this use of these models relies on the assumption that no group-level rater effects exist or, put differently, that the group-level data can be taken to represent error-free measurement. If this were true, in an Angoff setting, we would expect the results from the internal

frame of reference to correspond closely to the results of the same analysis from an external frame of reference constructed using the item response data from the original exam. As item response data was available for the current standard setting, it was possible to directly test this assumption. In the first part of the study, the estimates of the Angoff judges were used to construct an internal frame of reference, and the item difficulty parameter estimates from the original test administration were used to construct an external frame of reference. The ratings were analyzed separately in each frame using the MFRM and the results were compared. The values of the indices for leniency/severity, accuracy and centrality/extremity were found to differ across the two frames and it was thus concluded that the key assumption underlying use of the MFRM *did not hold in the present case*.

For leniency/severity, the differences were relatively minor, and only one additional judge was flagged in the external frame who was not flagged within the internal frame. Note that the comparison between the internal and external frames was indirect, as, of course, no external data was available for actual B1 students. (Making such a comparison would be possible using, e.g., the results from a standard setting using a different method. Such was beyond the scope of the present study.)

For inaccuracy, there were marked differences in the results and in the judges flagged for displaying the effect. Overall, the results from the external frame of reference revealed more rater effects than the internal frame of reference, and the correlational indices in the external frame suggested generally less positive or 'optimistic' views of judge performance.

For centrality/extremism, there were marked substantive differences, with a much larger number of judges were flagged within the external frame of reference. Indeed, judges who were flagged for showing an *extremity effect* in the internal frame *were actually found to show a centrality effect within the external frame.*

Overall, these results suggest that analysts need to carefully distinguish between normative (or 'rater consensus') situations and criterion-referenced situations. Without an external frame of reference, the use of latent-trait models for detecting rater effects can be used to make claims about the former type of situation, based on deviations from the ratings of the group of raters as a whole. However, the stronger claim of detecting deviations from 'error-free' measurement cannot be made without

considerable justification.

Group-level Effects. On a more positive note, the group-level effect indicators for centrality and inaccuracy within the internal frame of reference did suggest group-level effects, and analysis within the external frame confirmed that group-level effects did occur. This is of considerable importance, since in most rating situations, there is no data available from which an external frame of reference might be constructed. Indices capable of detecting group-level rater effects could be used to evaluate the assumption of the model *even in the absence of an external frame of reference*. If no group-level rater effects are found, it is more likely that the assumption is met and that the judge frame of reference does not differ significantly from an external 'error-free' frame of reference. On the other hand, finding group-level effects would serve to caution the analyst away from making criterion-referenced interpretations where only an internal frame of reference exists.

Assumptions of the Angoff Method

It was also found that the assumptions of the Angoff method appear to have been violated. Given previous findings in the literature, this was not particularly surprising. In terms of the first assumption (representation of the BPS), it was difficult to make clear statements about how accurately the cut score was located, for the simple reason that there is no external standard for comparison. In terms of the second assumption, in line with previous research, the present study showed that a high degree of inaccuracy was present in the first round of item estimates made by the judges. More optimistically, less inaccuracy was found in the listening test, suggesting that judges do become more accurate with experience. Finally, also in line with a number of earlier studies, raters were shown to have difficulty in quantifying their decisions using the probability scale, and almost all raters displayed a central tendency bias, compressing the internal scale.

It was also observed that judges who were non-native English speakers (as opposed to native English speakers) and judges with administrative roles (as opposed to judges with only teaching roles) were more severe, more accurate and less likely to

display centrality. The sample was too small to draw conclusions, but it is tempting to speculate that non-native speakers might be more able to perform the task required in an Angoff, since they have more direct experience with assessing the difficulty of target-language texts, compared to the native speakers, for whom it may represent more of an abstract exercise.

5.2 Implications and Suggestions

Use of the MFRM for Rater Detection

Incorporating Checks of the Assumption. A key finding of this study is that the assumption underlying use of the MFRM for detecting rater effects does not hold in all situations and that the model is *not* robust to violations of the assumption.

It is critical to again emphasize the distinction between normative and criterion-referenced settings. In normative-referenced settings, where the only concern is on achieving a set of consensus ratings and where 'error' is understood strictly in the sense of deviation from the consensus of the raters as a group, only an internal frame of reference is required, and the MFRM is a suitable tool. However, in criterion-referenced situations, where error is understood as deviation from the ratings that would be assigned based on explicit external criteria, use of the MFRM relies on the assumption that no group-level effects are present. Standard setting clearly represents a criterion-referenced setting, and thus evaluating this assumption is a necessary preliminary step.

Two possible methods exist for evaluating this assumption. Where data from an external frame of reference is available, the analysis can be conducted within both frames. If the results are the same, the argument that there are no group-level effects within the internal frame of reference is strengthened. However, if the results differ, further investigation is warranted. However, data from an external frame is unlikely to be present in most cases. Another important finding from this study is that group-level indicators for both inaccuracy and centrality *were* sensitive to group-level effects. In other words, indicators available within the internal frame of reference itself may

suffice to provide at least some indication of whether or not the MFRM is suitable for use within a criterion-referenced situation.

Efficacy of the Indicators. This study joins a growing body of research suggesting that fit statistics do not appear to be the best indicators for identifying inaccuracy or centrality, or for distinguishing between these two effects. Score-expected and point-measure correlations appear to be better suited for detecting inaccuracy. SR/ROR correlations may also be useful where only internal data is available. For centrality/ extremism, expected-residual and measure-residual correlations appeared to be most effective.

Implication for the Angoff Method

The present results concerning leniency/severity and inaccuracy were generally in line with earlier research. The findings concerning the centrality effect deserve further comment. As the present findings were consistent with previous findings, and as this problem is only slowly gaining more widespread attention despite its adverse impact on the outcome of a standard setting, this issue is discussed here at more length.

Taken together with previous reports in the standard setting literature, this study provides further reason for practitioners to be alert to the possible presence of the centrality or central tendency effect whenever the modified Angoff method is used. This bias likely poses a dual threat to the standard setting process, in that it not only distorts the cut score (and thus the pass/fail rate) but also distorts the measures used to evaluate the results of the standard setting. There is a clear and pressing need for further research into techniques for reducing or eliminating the centrality bias. Four concrete suggestions are suggested here.

1. Training. Training judges to avoid this effect is not currently a standard part of preparation for conducting a modified Angoff standard setting. If the conclusions presented here are accepted, it would certainly seem advisable to include such training in practice, and the results of training would clearly be of interest from a research perspective.

2. Procedural modification. A well-developed body of psychophsyics research on sequential effects suggest that centrality may result from the influence of the previous stimulus on the judgments made about the current stimulus (Cross, 1973; DeCarlo & Cross, 1990; Garner, 1953; Jesteadt, Luce, & Green, 1977; Ward, 1973). Teghtsoonian et al. (2008) were able to eliminate this sequentiality effect in one experiment by 'recalibrating' after each judgment to remove the influence of the previous judgment. Applied to the modified Angoff method, this suggests a procedural change. After each probability estimate, judges could be asked to recalibrate by, for example, by referring back to one or two items with their associated empirical p-values.

3. Scale equating. Obviously, where a centrality bias is present in one frame, its impact on the standard deviation will need to be taken into account if a common scale is to be maintained. Currently, however, this is not routinely performed as part of the Angoff procedure. Test equating is a frequently-used procedure for converting scores from two different tests (*specified frames of reference*) onto the same scale and is commonly used when educators want to directly compare the results of two tests (Wright & Stone, 1979). Within each separate frame of reference, the origin can be set at the mean item difficulty location and the difference in the means of the estimates for the common items or persons then serves as a constant used to adjust for the differences between the origins of the two scales. It could also occur that one of the tests was characterized by a greater amount of measurement error, resulting in weaker discrimination within that frame of reference, indicated by a larger standard deviation. This would indicate that the 'natural units' for the two frames differed, and a second constant could be used to account for this difference in unit size between the two frames. Sometimes referred to as 'mean/sigma equating' within the context of test equating (Embretson & Reise, 2000; Skaggs & Wolfe, 2009), the maintenance of a common, arbitrary unit is important in other measurement settings as well, and the *extended frame of reference model* (EFRM) provides a general form in which person and item locations from different *specified frames of reference* can be located in terms of a common scale, with a common

arbitrary unit and a common origin (Humphry, 2005; 2011; Humphry & Andrich, 2008).

An Angoff standard setting also represents a situation in which measures from two frames of reference need to be placed on a common scale, although this has only rarely been explicitly addressed (Heldsinger, 2006; Heldsinger & Humphry, 2006). In an Angoff setting, the test for which the standard is set is analagous to the 'anchor test' in an equating situation. The origin of the test on the logit scale is the common origin *c*, and the natural unit within the test frame of references is the common unit *u*. The problem then becomes one of expressing the cut score locations from the Angoff judges in terms of the common origin (*c*) and common unit (*u*). This can be accomplished with the following conversions:

$$\beta_n^T = \rho_T(\beta_n^A - c_T^A); \text{ and} \tag{5.1}$$

$$\delta_i^T = \rho_T(\delta_i^A - c_T^A); \text{ and} \tag{5.2}$$

$$\rho_T = \frac{\sigma^T}{\sigma^A} \tag{5.3}$$

where the superscripts 'T' and 'A' refer to the test and Angoff judge frames of reference respectively, c_T^A is the magnitude of the difference between the common origin from the test and the origin of the Angoff frame of reference, and ρ_T is a frame discrimination parameter used to adjust for differences in unit size between the test and Angoff frames of reference. The frame discrimination parameter, ρ_T, is defined by using the standard deviations within each frame as the natural units.

The present study used an alternative method for maintaining a common scale, using the Rasch model to anchor item difficulty measures to their values within an external frame. It would clearly be of interest to compare results from the two methods, and to attempt to determine which method is more suitable.

4. Use of alternative standard setting methods. Finally, there are standard setting methods, such as the bookmark method, in which the scale from the original frame of reference is preserved, so that the judges do not need to construct a new scale through their estimates (Peterson et al., 2011).

Considering using such methods instead of the Angoff method would avoid the centrality bias altogether.

5.3 Limitations of the Present Study

Four clear limitations of this study stand out enough to deserve being stated explicitly. First, the study drew primarily on data from only the *first* round of a modified Angoff standard setting, before empirical feedback data was introduced and before judges had a chance to discuss and reconsider their decisions. Second, only one possible configuration of the MFRM was used - a binomial model with the original estimates recoded into 11 categories. It is possible that different models and a different number of categories would have yielded different results. Third, this study did not seek to differentiate between random and systematic sources of inaccuracy. Pursuing the sources of inaccuracy is a task of obvious importance. Finally, this study made use of empirical data from two meetings with a limited sample of items and panelists. Any conclusions should be drawn with these limitations in mind.

5.4 Future Research Directions

This study highlights certain areas in which much important work remains to be done. First, for use of the MFRM for criterion-related purposes such as standard setting, it seems clear that it is vital to evaluate the plausibility of the 'no group-level rater effects' assumption *before* using the model. In the absence of data from an external framework, group-level indicators may be the only means of assessing this assumption. Therefore, further research on such indicators is clearly warranted.

Further, this study is consistent with a number of earlier studies which suggest that a central tendency bias may be inherent in the way the Angoff procedure is structured. This suspicion finds a theoretical basis is research within the psychophysics tradition. Given the widespread use of the Angoff method in high stakes situations, more research into how this bias can be avoided or adjusted for is a clear priority.

Also, further work, perhaps in the form of simulation studies, is warranted to provide a more refined understanding of which indices are likely to be most sensitive to particular rater effects.

Finally, in reference to judge characteristics, it was suggested here that non-native speakers of a language being tested may be better equipped to perform the tasks required in an Angoff setting. This could be further researched through a study designed for the purpose, involving having NS and NNS participants estimate item difficulty for a given student population for which item response data is available.

REFERENCES

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington, DC: American Council on Education.

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

Brennan, R.L. & Lockwood, R.E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. *Applied Psychological Measurement, 4,* 219-240.

Bourque, M.L. (2000, April). Setting student performance standards: The role of achievement level descriptions in the standard setting process. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Busch, J.C., & Jaeger, R.M. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher Examinations. *Journal of Educational Measurement, 27*, 145-163.

Chang, L. (1999) Judgmental item analysis of the Nedelsky and Angoff standard-setting methods. *Applied Measurement in Education, 12,* 151-165.

Chang, L., Dzuiban, C.D., & Olson, A.H. (1996). Does a standard reflect minimal competency of examinees or judge competency? *Applied Measurement in Education, 9,* 161-173.

Cizek, G.J., & Bunch, M.B. (2007). Standard setting: A guide to establishing and evaluating performance standards on tests. Thousand Oaks, CA: Sage.

Clauser, B.E., Harik, P., Margolis, M.J., McManus, I.C., Mollon, J., Chis, L., & Williams, S. (2009). An empirical examination of the impact of group discussion

and examinee performance information on judgments made in the Angoff standard-setting procedure. *Applied Measurement in Education, 22*, 1-21.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

Council of Europe. (2009). Manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment. Strasbourg, France: Council of Europe/Language Policy Division.

Cross, D.V. (1973). Sequential dependencies and regression in psychophysical judgment. *Perception & Psychophysics, 14,* 547-552.

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

DeCarlo, L.T., & Cross, D.V. (1990). Sequential effects in magnitude scaling: Models and theory. *Journal of Experimental Psychology: General, 119,* 375-396.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, *2*, 197–221.

Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (Section H).* Strasbourg, France: Council of Europe/Language Policy Division.

Egan, K.L., Ferrara, S., Schneider, M.C., & Barton, K. (2009). Writing Performance Level Descriptors and Setting Performance Standards for Assessments of Modified Achievement Standards: The Role of Innovation and Importance of Following Conventional Practice', *Peabody Journal of Education*, *84*(4), 552-577.

Engelhard, G., Jr. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education, 5,* 171-191.

Engelhard, G., Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational*

Measurement, 31, 93-112.

Engelhard, G., Jr. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement, 33,* 56-70.

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

Engelhard, G., Jr. (2009). Evaluating the judgments of standard-setting panelists using Rasch measurement theory. In *Criterion referenced testing: Practice analysis to score reporting using Rasch measurement models* (pp. 312-346). Maple Grove, Minnesota: JAM Press.

Engelhard, G., Jr. (2011). Evaluating the bookmark judgments of standard-setting panelists. *Educational and Psychological Measurement, 71*(6), 909-924.

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

Englehard, G., Jr., & Cramer, S. (1997). Using Rasch Measurement to evaluate the ratings of standard-setting judges. In M. Wilson, G. Engelhard, and K. Draney. (Eds.). *Objective measurement: theory into practice* (Vol. 4, pp. 97-112). Norwood, NJ: Ablex.

Engelhard, G., & Gordon, B. (2000). Setting and evaluating performance standards for high stakes writing assessments. In M. Wilson & G. Engelhard (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp. 3–14). Stamford, CT: Ablex.

Englehard, G., Jr., and Stone, G.E. (1998). Evaluating the quality of ratings obtained from standard-setting judges. *Educational and Psychological Measurement, 58*(2), 179-196.

Fehrmann, M.L., Woehr, D.J., & Arthur, W., Jr. (1991). The Angoff cutoff score method: The impact of frame of reference rater training. *Educational and Psychological Measurement, 51*, 857-872.

Ferdous, A.A., & Plake, B.S. (2005). Understanding the factors that influence decisions of panelists in a standard setting study. *Applied Measurement in Education, 18*(3), 257-267.

Garner, W.R. (1953). An informational analysis of absolute judgments of loudness.

Journal of Experimental Psychology, 46, 373-380.

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

Giraud, G., Impara, J.C., & Plake, B.S. (2000, April). *A qualitative examination of teachers' conception of the just competent examinee in Angoff workshops.* Paper presented at the meeting of the American Educational Research Association, New orleans, LA.

Giraud, G., Impara, J.S., & Plake, B.S. (2005). Teachers' Conceptions of the Target Examinee in Angoff Standard Setting. *Applied Measurement in Education, 18*(3), 223-232.

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

Hamberlin, M.K. (1992). *Influence of item response theory and type of judge on a standard set using the iterative Angoff standard setting method. Unpublished doctoral dissertation, University of North Texas, Denton, TX.*

Heldsinger, S. (2006). *Accounting for unit of scale in standard setting methodologies* (Doctoral dissertation, Murdoch University, Perth, Australia). Retrieved from http://researchrepository.murdoch.edu.au/72/

Heldsinger, S., & Humphry, S. (2006). *Maintaining consistent metrics in standard setting.* Unpublished manuscript, Murdoch University, Perth, Australia.

Humphry, S. (2005). *Maintaining a Common Arbitrary Unit in Social Measurement* (Doctoral dissertation, Murdoch University, Perth, Australia). Retrieved from http://wwwlib.murdoch.edu.au/adt/browse/view/adt-MU20050830.95143

Humphry, S. (2011). The role of the unit in physics and psychometrics. *Measurement: Interdisciplinary Research & Perspective. 9*(1), 1-24.

Humphry, S., & Andrich, D. (2008). Understanding the Unit in the Rasch Model. *Journal of Applied Measurement, 9*(3), 249-264.

Hollingworth, H.L. (1910). The central tendency of judgment. *The Journal of Philosophy, Psychology and Scientific Methods, 7*(17), 461-469.

Hurtz, G.M., & Jones, J.P. (2009) Innovations in measuring rater accuracy in standard setting: Assessing 'fit' to item characteristic curves. *Applied Measurement in*

Education, 22, 120-143.

Impara, J.C. (1997, October). *Setting standards using Angoff's method: Does the method meet the standard?* Paper presented to the Midwestern Educational Research Association, Chicago.

Impara, J.C., Giraud, G., & Plake, B.S. (2000, April). *The influence of providing target group descriptors when setting a passing score.* Paper presented at the meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED445013).

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

Jaeger, R.M. (1989). Certification of student competence. In R.L. Linn (Ed.), *Educational Measurement* (3rd ed.). Washington, DC: American Council on Education.

Jaeger, R.M. (1991). Selection of judges for standard-setting. *Educational Measurement: Issues and Practice, 10,* 3-6.

Jesteadt, W., Luce, R. D., & Green, D. M. (1977). Sequential effects in judgments of loudness. *Journal of Experimental Psychology: Human Perception & Performance, 3,* 92–104.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, *64*(3), 425-461.

Kim, S. C., & Wilson, M. (2009). A Comparative Analysis of the Ratings in Performance Assessment Using Generalizability Theory and The Many-Facet Rasch Model. *Journal of Applied Measurement*, *10*(4), 40-423.

Lewis, D.M. & Green, D.R. (1997). *The validity of performance level descriptors.* Paper presented at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

Linacre, J.M. (1989). *Many-Facet Rasch Measurement*. Chicago: MESA Press.

Linacre, J.M. (2000). Using Rasch fit statistics to rescale linear measures and anchor values. *Rasch Measurement Transactions, 14*(2), 750. Retrieved from http:// www.rasch.org/rmt/rmt142n.htm

Linacre, J.M. (2009). Facets Rasch measurement computer program (Version 3.68.0).

Chicago: Winsteps.com.

Linn & Gronlund, (2000). *Measurement and Assessment in Teaching (Eighth Edition)*. Des Moines: Prentice-Hall.

Lorge, I., & Kruglov, L.K. (1953). The improvement of the estimates of test difficulty. *Educational and Psychological Measurement, 13*, 34-46.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*, 54–71.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*(2), 158-180.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Maurer, T.J., Alexander, R.A., Callahan, C.M., Bailey, J.J., & Dabrot, F.H. (1991). Methodological and psychometric issues in setting cutoff scores using the Angoff method. *Personnel Psychology, 44,* 235-262.

Mercado, R. L., & Egan, K. L. (2005). *Performance level descriptors*. Paper presented at the National Council on Measurement in Education, Montreal, Quebec, Canada.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

McGinty, D. (2005). Illuminating the 'Black Box' of Standard Setting: An exploratory qualitative study. *Applied Measurement in Education, 18*(3), 269-287.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*, 386–422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*, 189–227.

Newcomb, T. (1931). An experiment designed to test the validity of a rating technique. *Journal of Educational Psychology, 22*(4). 279-289.

Norcini, J.J., Shea, J.A., & Kanya, D.T. (1988). The effects of various factors on standard setting. *Journal of Educational Measurement, 25,* 57-65.

Noor, Lide Binti Abu Kassim. (2007). *Using the Rasch measurement model for standard setting of the English Language Placement Test at the IIUM* (Unpublished doctoral dissertation), Universiti Sains Malaysia, Pulau Pinang, Malaysia.

Papageorgiou, S. (2010). Investigating the decision-making process of standard setting participants. *Language Testing, 27*(2), 261-282.

Peterson, C.H., Schulz, E.M., & Engelhard, G., Jr. (2011). Reliability and validity of bookmark-based methods for standard setting: Comparisons to Angoff-based methods in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice, 30*(2), 3-14.

Plake, B.S., & Impara, J.C. (2001). Ability of panelists to estimate item performance for a target group of candidates: An issue in judgmental standard setting. *Educational Assessment, 7,* 87-97.

Plake, B.S., Impara, J.C., & Irwin, P. (2000). Consistency of Angoff-based predictions of item performance: Evidence of technial quality of results from the Angoff standard setting method. *Journal of Educational Measurement, 37,* 347-355.

Plake, B.S., Impara, J.C., & Potenza, M.T. (1994). Content specificity of expert judgements in a standard setting study. *Journal of Educational Measurement, 31,* 339-347.

Poulton, E.C. (1979). Models for biases in judging sensory magnitude. *Psychological Bulletin, 86*(4), 777-803.

Pula, J.J., & Huot, B.A. (1993). A model of background influences on holistic raters. In M.M. Williamson & B.A. Huot (Eds.), *Validuting holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 237-265). Cresskill, NJ: Hampton Press.

Rasch, G. (1977). On Specific Objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *The Danish Yearbook of Philosophy, 14,* 58-93. Retrieved from http://www.rasch.org/memo18.htm

Reckase, M.D. (2006). A conceptual framework for a psychometric theory for standard setting with examples of its use for evaluating the functioning of two standard setting methods. *Educational Measurement: Issues and Practice, 25*(2), 4-18.

Reid, J.B. (1985, April). *Establishing upper limits for item rtings for the Angoff method: Are resulting standards more 'realistic'?* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago.

Ricker, K.L. (2006). Setting cut-scores: A critical review of the Angoff and Modified Angoff Methods. *Alberta Journal of Educational Research, 52*(1), 53-64.

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.

Schulz, E.M. (2006). Commentary: A response to Reckase's conceptual framework and examples for evaluating standard setting. *Educational Measurement: Issues and Practice, 25*(3), 4-13.

Scullen, S.E., Mount, M.K., & Goff, M. (2000). Understanding the Latent Structure of Job Performance Ratings. *Journal of Applied Psychology, 85*(6), 956-970.

Skorupski, W.P., & Hambleton, R.K. (2005). What are panelists thinking when they participate in standard-setting studies? *Applied Measurement in Education, 18*(3), 233-256.

Shepard, L.A. (1994, October). *Implications for standard setting of the NAE evaluation of NAEP achievement levels.* Paper presented at the Joint Conference on Standard Setting for Large Scale Assessments, National Assessment Governing Board, National Center for Educational Statistics, Washington, DC.

Shepard, L., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement tests.* Stanford, CA: National Academy of Education.

Stemler, S.E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, *9*(4). Retrieved from http://PAREonline.net/getvn.asp?v=9&n=4

Stemler, S. E., & Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 29–49). Los Angeles: Sage.

Stevens, S.S., & Greenbaum, H.B. (1966). Regression effect in psychophysical judgment. *Perception & Psychophysics, 1*, 439-446.

Teghtsoonian, R. (1973). Range effects in psychophysical scaling and a revision of Stevens's law. *American Journal of Psychology, 86,* 3–27.

Teghtsoonian, R., & Teghtsoonian, M. (1978). Range and regression effects in magnitude scaling. *Perception & Psychophysics, 24*, 305–314.

Teghtsoonian, M., Teghtsoonian, R., & DeCarlo, L.T. (2008). The influence of trial-to-trial recalibration on sequential effects in cross-modality matching. *Psychological Research*, *72,* 115-122.

van der Linden, W.J. (1982). A latent trait method for determining intrajudge consistency in the Angoff and Nedelsky techniques of standard setting. *Journal of Educational Measurement, 19,* 295-308.

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., Van der Stegg, A.F.W., Scherpbier, A.J.J.A., Muijtjens, A.M.M., Verwijnen, G.M., & Van der Vleuten, C.P.M. (1999). Reliability and credibility of an Angoff standard setting procedure in progress testing using recent graduates as judges. *Medical Education, 33,* 832-837.

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 compareted to recently graduated students. *Medical Education, 36,* 860-867.

Ward, L. M. (1973). Repeated magnitude estimations with a variable standard: Sequential effects and other properties. *Perception & Psychophysics, 14,* 193–200.

Weir, J.C. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, *22*(3), 1-20.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, *10*, 305–335.

Wilson, M., & Case, H. (2006). An examination of variation in rater severity over time: A study in rater drift. In M. Wilson & G. Engelhard, Eds., *Objective measurement: theory into practice*.

Wolfe, E.W. (2004). Identifying rater effects using latent trait models. *Psychology Science, 46*, 35-51.

Wolfe, E.W., Chiu, C.W.T., & Myford, C.M. (2000). Detecting rater effects with a multi-faceted Rasch rating scale model. In M.Wilson & G. Engelhard (Eds.), *Objective measurement: Theory into practice* (Vol. 5, pp. 147-164). Stamford, CT: Ablex.

Wolfe, E.W., & McVay, A. (2010). Rater effects as a function of rater training context. Retrieved from http://www.pearsonassessments.com/NR/rdonlyres/6435A0AF-0C12-46F7-812E-908CBB7ADDFF/0/RaterEffects_101510.pdf

Wolfe, E.W., & McVay, A. (2011, April). Application of latent trait models to identifying substantively interesting raters. Presented at the Annual Conference of the American Educational Research Association, New Orleans. Retrieved from http://www.pearsonassessments.com/hai/images/PDF/AERA_Application_Latent_Trait_Models.pdf

Wright, B.D., & Linacre, J. M. (1994). Reasonable mean-square fit values. Rasch Measurement Transactions, 8(3), 370.

Wright, B.D., & Masters, G. N. (1982). Rating scale analysis: Rasch measurement. Chicago: MESA.

Wright, B.D., & Stone, M.H. (1979). *Best test design*. Chicago: MESA.

Yue, Xiaohui. (2011). *Detecting rater centrality effect using simulation methods and Rasch measurement analysis* (Doctoral dissertation, Virginia Polytechnic Institute and State University). Retrieved from http://scholar.lib.vt.edu/theses/available/etd-07272011-104720/unrestricted/Yue_X_D_2011.pdf

Appendix A-1. Item Quality Statistics from Original Administration of Test, Reading

Item	Measure (logits)	InFit Mn Sq	OutFit Mn Sq	Item Pt Bsrl	Item	Measure (logits)	InFit Mn Sq	OutFit Mn Sq	Item Pt Bsrl
iR1	1.04	1.02	1.08	0.34	iR26	-2.00	0.95	0.75	0.20
iR2	-0.31	1.02	0.98	0.34	iR27	-0.29	1.00	1.00	0.37
iR3	-1.52	0.88	0.67	0.44	iR28	-1.92	0.84	0.70	0.44
iR4	-1.32	0.96	0.86	0.24	iR29	-0.99	0.94	0.84	0.40
iR5	0.35	0.91	0.84	0.39	iR30	-1.15	0.96	0.91	0.38
iR6	-0.61	0.94	0.85	0.42	iR31	-0.24	0.96	0.93	0.41
iR7	-0.63	1.01	0.96	0.34	iR32	-0.7	0.933	0.8494	0.4307
iR8	1.20	1.01	1.03	0.26	iR33	-1.19	0.9361	0.9058	0.4046
iR9	0.32	1.03	1.05	0.22	iR34	0.9	1.1199	1.2024	0.2758
iR10	0.72	1.02	1.01	0.25	iR35	-0.4	0.9387	0.9007	0.4335
iR11	-2.01	0.92	0.75	0.31	iR36	-1.03	0.9038	0.7603	0.333
iR12	-0.31	1.13	1.24	0.15	iR37	-0.06	0.9898	0.9709	0.2683
iR13	0.91	1.04	1.07	0.30	iR38	-1.15	0.9332	0.9168	0.2978
iR14	1.11	1.01	1.05	0.31	iR39	0.81	0.9374	0.9281	0.3145
iR15	-0.43	0.94	0.87	0.41	iR40	-0.51	0.8674	0.7653	0.3598
iR16	-0.19	0.94	0.92	0.36					
iR17	1.18	1.07	1.10	0.21					
iR18	0.86	1.02	1.02	0.28					
iR19	1.62	1.17	1.29	0.08					
iR20	1.55	1.05	1.08	0.23					
iR21	0.72	1.14	1.17	0.12					
iR22	1.48	0.99	1.00	0.27					
iR23	0.10	1.08	1.11	0.15					
iR24	2.28	1.03	1.13	0.12					
iR25	1.85	1.00	1.04	0.18					

Appendix A-2. Item Quality Statistics from Original Administration of Test, Listening

Item	Measure (logits)	InFit Mn Sq	OutFit Mn Sq	Item Pt Bsrl	Item	Measure (logits)	InFit Mn Sq	OutFit Mn Sq	Item Pt Bsrl
iL1	0.38	1.04	1.05	0.30	iL26	0.11	0.92	0.86	0.42
iL2	-0.77	0.93	0.88	0.37	iL27	1.40	0.93	0.94	0.37
iL3	-0.75	0.96	0.90	0.35	iL28	0.92	1.00	0.99	0.33
iL4	0.76	1.02	1.05	0.32	iL29	-0.75	1.00	0.95	0.32
iL5	-0.20	0.89	0.82	0.45	iL30	0.76	0.90	0.86	0.45
iL6	0.11	0.93	0.91	0.41	iL31	1.15	1.07	1.11	0.26
iL7	-1.47	0.92	0.76	0.36	iL32	-1.33	0.96	0.91	0.31
iL8	-0.71	0.93	0.85	0.39	iL33	0.65	1.05	1.10	0.31
iL9	-0.78	0.97	0.96	0.33	iL34	-0.2	1.03	1.03	0.32
iL10	-1.12	0.97	0.86	0.32	iL35	1.43	0.96	0.98	0.38
iL11	0.89	0.97	0.98	0.38	iL36	0.48	0.96	0.93	0.37
iL12	-0.55	0.93	0.82	0.40	iL37	-1.84	0.94	0.72	0.31
iL13	0.16	1.00	1.00	0.32	iL38	-1.06	0.93	0.81	0.37
iL14	2.06	1.15	1.39	0.11	iL39	0.14	0.90	0.82	0.44
iL15	-0.21	0.97	0.97	0.34	iL40	-0.14	0.96	0.94	0.36
iL16	1.30	0.94	0.95	0.37					
iL17	0.02	1.04	1.07	0.29					
iL18	0.67	1.05	1.07	0.28					
iL19	0.83	0.94	0.93	0.39					
iL20	0.88	1.09	1.20	0.23					
iL21	-1.52	0.91	0.81	0.37					
iL22	-0.88	0.94	0.94	0.37					
iL23	-1.11	0.92	0.83	0.38					
iL24	-0.28	0.88	0.81	0.46					
iL25	0.57	1.09	1.09	0.23					

Appendix B-1. CEFR Scales Used to Provide Performance Level Descriptors (PLDs)

Scale No.	Scale Title (CEFR page)
LISTENING	
1	Common Reference Levels: global scale (p. 24)
13	OVERALL LISTENING COMPREHENSION (p. 66)
14	UNDERSTANDING CONVERSATION BETWEEN NATIVE SPEAKERS (p. 66)
15	LISTENING AS A MEMBER OF A LIVE AUDIENCE (p. 67)
16	LISTENING TO ANNOUNCEMENTS AND INSTRUCTIONS (p. 67)
17	LISTENING TO AUDIO MEDIA AND RECORDINGS (p. 68)
24	IDENTIFYING CUES AND INFERRING (p. 72)
26	UNDERSTANDING A NATIVE SPEAKER INTERLOCUTOR (p. 75)
27	CONVERSATION (P. 76)
28	INFORMAL DISCUSSION (WITH FRIENDS) (P. 77)
29	FORMAL DISCUSSION AND MEETINGS (P. 78)
30	GOAL-ORIENTED CO-OPERATION (P. 79)
31	TRANSACTION TO OBTAIN GOODS AND SERVICES (P. 80)
32	INFORMATION EXCHANGE (P. 81)
READING	
1	Common Reference Levels: global scale (p. 24)
2	Common Reference Levels: self-assessment grid (p. 26)
20	OVERALL READING COMPREHENSION (p. 69)
21	READING CORRESPONDENCE (p. 69)
22	READING FOR ORIENTATION (p. 70)
23	READING FOR INFORMATION AND ARGUMENT (p. 70)
24	READING INSTRUCTIONS (p. 71)
26	IDENTIFYING CUES AND INFERRING (Spoken & Written) (P. 72)

Appendix B-2. CEFR Global Scale (CEFR, p. 24)

CEFR Level	**Common Reference Levels: global scale (CEFR, p. 24)**
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 Proficient 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.

Appendix C. Angoff Judge Response Form

(Note: Only the first page is reproduced below. The pages for the remaining 32 items are identical.)

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 D-1. Correlational Indices of Inaccuracy for Reading, Internal v. External

	Correlation of Scores with p-values		Point-Biserial Correlation		Score-Expected Correlation		Point-Measure Correlation	
Judges	INT.	EXT.	INT.	EXT.	INT.	EXT.	INT.	EXT.
J01		0.33*	0.59		0.64	.32*	0.64	.381*
J02		0.54	0.79		0.83	0.54	0.83	0.56
J03		0.57	0.56		0.64	0.56	0.63	0.55
J04		0.38*	0.38*		0.44	0.37*	0.45	0.37*
J05		0.43	0.44		0.49	0.40	0.49	0.39*
J06		0.21**	0.43		0.50	0.24**	0.50	0.27**
J07		-0.04**	0.25**		0.34*	-0.03**	0.35*	-0.02**
J08		0.31**	0.35*		0.39*	0.29**	0.40	0.30**
J09		0.43	0.59		0.64	0.44	0.63	0.44
J10		0.06**	0.24**		0.3**	0.04**	0.31**	0.06**
J11		0.12**	0.40*		0.46	0.10**	0.47	0.10**
J12		0.30**	0.39*		0.46	0.29**	0.45	0.29**
J13		0.63	0.61		0.65	0.64	0.64	0.63
J14		0.54	0.76		0.80	0.55	0.79	0.56
J15		0.41	0.77		0.82	0.43	0.82	0.44
J16		0.55	0.79		0.83	0.55	0.83	0.55
J17		0.38*	0.71		0.76	0.39*	0.76	0.40*
J18		0.65	0.81		0.85	0.68	0.85	0.68
Accurate		9	12		15	9	16	8
*Marginal		3	4		2	3	1	4
**Inaccurate		6	2		1	6	1	6

*p > .01; **p > .05 (Note that asterisks indicate low or non-significant correlations.)

Appendix D-2. Judge Fit Indices of Inaccuracy for Reading, Internal v. External

	InFit Mean Square		InFit Standardized		OutFit Mean Square		OutFit Standardized		SD Residuals	
Judges	INT.	EXT.	INT.	EXT.	INT.	EXT.	INT.	EXT.	INT.	EXT.
J01	0.56	1.87*	-2.28	3.06*	0.51	1.66*	-2.64	2.52*	0.94	1.47*
J02	0.39	1.55*	-3.65	2.19*	0.38	1.50*	-3.75	2.05*	0.90	1.50*
J03	1.11	1.89*	0.55	3.20*	1.00	1.69*	0.09	2.65*	1.45*	1.60*
J04	0.63	1.77*	-1.88	2.86*	0.60	1.85*	-2.07	3.14*	1.11	1.56*
J05	0.42	1.64*	-3.37	2.51*	0.42	1.83*	-3.38	3.12*	0.95	1.57*
J06	0.70	2.36*	-1.46	4.53*	0.67	2.33*	-1.65	4.51*	1.20	1.84*
J07	1.01	3.58*	0.14	7.29*	1.02	3.74*	0.16	7.64*	1.50*	2.36*
J08	0.56	1.73*	-2.32	2.70*	0.53	1.83*	-2.51	3.09*	0.99	1.48*
J09	0.43	1.66*	-3.26	2.55*	0.42	1.71*	-3.36	2.73*	0.95	1.55*
J10	0.55	2.33*	-2.39	4.48*	0.54	2.50*	-2.46	4.94*	1.08	1.86*
J11	0.69	2.78*	-1.54	5.57*	0.67	3.00*	-1.65	6.10*	1.22	2.05*
J12	0.61	2.03*	-2.02	3.65*	0.62	2.14*	-1.94	3.99*	1.12	1.71*
J13	0.35	1.05	-3.99	0.27	0.33	1.08	-4.17	0.44	0.82	1.20
J14	0.57	1.64*	-2.30	2.49*	0.53	1.56*	-2.53	2.24*	1.06	1.52*
J15	0.81	2.65*	-0.89	5.28*	0.80	2.56*	-0.92	5.10*	1.33*	2.01*
J16	0.81	2.07*	-0.88	3.77*	0.78	2.03*	-1.04	3.71*	1.30*	1.74*
J17	0.57	2.28*	-2.29	4.38*	0.56	2.29*	-2.36	4.43*	1.12	1.87*
J18	0.78	1.55*	-1.01	2.20*	0.78	1.47*	-1.05	1.95*	1.30*	1.53*
Accurate	18	1	18	1	18	1	18	1	13	1
*Inaccurate	0	17	0	17	0	17	0	17	5	17

*Flagged for inaccuracy (Fit Mean Square > 1.40; Fit Standardized > 2.0, SD Residuals > 1.25)

Appendix D-3. Correlational Indices of Inaccuracy for Listening, Internal v. External

	Correlation of Scores with p-values		Point-Biserial Correlation		Score-Expected Correlation		Point-Measure Correlation	
Judges	INT.	EXT.	INT.	EXT.	INT.	EXT.	INT.	EXT.
J01		0.45	0.7		0.74	0.48	0.75	0.50
J02		0.72	0.66		0.71	0.71	0.71	0.71
J03		.433*	0.58		0.62	0.48	0.64	0.49
J04		0.56	0.69		0.74	0.56	0.74	0.56
J05		.368*	0.6		0.65	0.39*	0.65	0.40*
J06		0.47	0.64		0.68	0.47	0.69	0.48
J07		0.29**	0.34*		0.41	0.23**	0.40*	0.23**
J08		0.42	0.7		0.74	0.41	0.73	0.42
J09		0.55	0.71		0.75	0.55	0.74	0.55
J10		0.53	0.48		0.57	0.61	0.58	0.61
J11		-0.06**	0.02**		0.12**	-0.06**	0.13**	-0.04**
J12		0.69	0.67		0.72	0.67	0.72	0.67
J13		0.66	0.62		0.67	0.68	0.66	0.67
J14		0.57	0.59		0.66	0.56	0.62	0.54
J15		0.65	0.69		0.74	0.67	0.73	0.66
J16		0.60	0.73		0.78	0.60	0.75	0.58
J17		.395*	0.77		0.80	0.37*	0.80	0.39*
J18		0.46	0.72		0.76	0.41*	0.75	0.42
Accurate		13	16		17	13	16	14
*Marginal		3	1		0	3	1	2
**Inaccurate		2	1		1	2	1	2

*$p > .01$; **$p > .05$ (Note that asterisks indicate low or non-significant correlations.)

Appendix D-4. Judge Fit Indices of Inaccuracy for Listening, Internal v. External

	InFit Mean Square		InFit Standardized		OutFit Mean Square		OutFit Standardized		SD Residuals	
Judges	INT.	EXT.	INT.	EXT.	INT.	EXT.	INT.	EXT.	INT.	EXT.
J01	0.29	1.06	-4.49	0.35	0.28	1.05	-4.67	0.29	0.73	1.21
J02	0.30	0.61	-4.43	-1.95	0.29	0.59	-4.52	-2.11	0.73	0.92
J03	0.28	0.89	-4.53	-0.42	0.28	0.94	-4.61	-0.19	0.66	1.04
J04	0.51	1.14	-2.64	0.67	0.49	1.08	-2.79	0.42	0.97	1.27*
J05	0.31	1.20	-4.34	0.93	0.31	1.22	-4.31	1.01	0.75	1.30*
J06	0.22	1.02	-5.33	0.18	0.21	1.07	-5.47	0.37	0.65	1.23
J07	0.47	1.57*	-3.00	2.29*	0.47	1.70*	-3.01	2.72*	0.98	1.57*
J08	0.25	1.12	-4.92	0.58	0.24	1.10	-5.07	0.51	0.66	1.22
J09	0.28	0.98	-4.73	-0.01	0.26	1.00	-4.88	0.09	0.74	1.22
J10	0.57	0.99	-2.28	0.01	0.57	1.03	-2.28	0.23	1.11	1.27*
J11	0.89	2.37*	-0.44	4.58*	0.88	2.39*	-0.51	4.67*	1.32*	1.89*
J12	0.28	0.71	-4.66	-1.37	0.28	0.74	-4.67	-1.22	0.73	1.03
J13	0.23	0.69	-5.17	-1.51	0.23	0.74	-5.20	-1.25	0.67	1.00
J14	0.50	1.01	-2.76	0.11	0.48	1.00	-2.86	0.08	0.93	1.17
J15	0.36	0.76	-3.88	-1.09	0.34	0.75	-4.04	-1.18	0.83	1.05
J16	0.34	0.87	-4.05	-0.51	0.31	0.91	-4.32	-0.36	0.77	1.09
J17	0.19	1.25	-5.81	1.11	0.18	1.25	-5.87	1.13	0.59	1.35*
J18	0.20	1.06	-5.58	0.35	0.18	1.06	-5.78	0.33	0.58	1.18
Accurate	18	16	18	16	18	16	18	16	17	12
*Inaccurate	0	2	0	2	0	2	0	2	1	6

*Flagged for inaccuracy (Fit Mean Square > 1.40; Fit Standardized > 2.0, SD Residuals > 1.25)

Appendix D-5. Indices of Centrality/Extremity for Reading, Internal v. External

		$r_{exp,res}$		$r_{meas,res}$	
Judge	Raw Score Standard Deviation	INTERNAL	EXTERNAL	INTERNAL	EXTERNAL
J01	11.9	0.08	-0.62*	0.09	-0.56*
J02	15.4	0.36*	-0.55*	0.36*	-0.54*
J03	17.8	0.24	-0.30	0.22	-0.31
J04	11.2	-0.31	-0.72*	-0.30	-0.71*
J05	9.6	-0.51*	-0.84*	-0.50*	-0.84*
J06	13.7	-0.19	-0.70*	-0.18	-0.68*
J07	14.7	-0.31	-0.76*	-0.30	-0.76*
J08	9.8	-0.39*	-0.77*	-0.37*	-0.75*
J09	12.2	-0.14	-0.71*	-0.14	-0.71*
J10	8.3	-0.62*	-0.88*	-0.60*	-0.87*
J11	13.2	-0.28	-0.77*	-0.28	-0.76*
J12	11.9	-0.32*	-0.75*	-0.33*	-0.74*
J13	10.7	-0.20	-0.73*	-0.21	-0.73*
J14	16.3	0.39*	-0.45*	0.38*	-0.45*
J15	19.4	0.51*	-0.43*	0.51*	-0.42*
J16	19.7	0.55*	-0.33*	0.54*	-0.33*
J17	16.3	0.27	-0.58*	0.27	-0.57*
J18	20.1	0.59*	-0.23	0.59*	-0.22
*Centrality		4	16	4	16
No Effect		9	2	9	2
*Extremity		5	0	5	0

*$p < .05$

Appendix D-6. Judge Fit Indices of Centrality/Extremity for Reading, Internal v. External

		InFit Mean Square		SD Residuals	
Judges	Raw Score Standard Deviation	INT.	EXT.	INT.	EXT.
J01	11.9	0.56*	1.87*	0.94	1.47*
J02	15.4	0.39*	1.55*	0.90	1.50*
J03	17.8	1.11	1.89*	1.45*	1.60*
J04	11.2	0.63	1.77*	1.11	1.56*
J05	9.6	0.42*	1.64*	0.95	1.57*
J06	13.7	0.70	2.36*	1.20	1.84*
J07	14.7	1.01	3.58*	1.50*	2.36*
J08	9.8	0.56*	1.73*	0.99	1.48*
J09	12.2	0.43*	1.66*	0.95	1.55*
J10	8.3	0.55*	2.33*	1.08	1.86*
J11	13.2	0.69	2.78*	1.22	2.05*
J12	11.9	0.61	2.03*	1.12	1.71*
J13	10.7	0.35*	1.05	0.82	1.20
J14	16.3	0.57	1.64*	1.06	1.52*
J15	19.4	0.81	2.65*	1.33*	2.01*
J16	19.7	0.81	2.07*	1.30*	1.74*
J17	16.3	0.57	2.28*	1.12	1.87*
J18	20.1	0.78	1.55*	1.30*	1.53*
*Overfit/ Centrality		7	0	0	0
Acceptable		11	1	13	1
*Underfit/ Extremity		0	17	5	17

*For Infit Mean Square: Overfit = fit < 0.60; underfit = fit > 1.4.
For SD Residuals: Centrality = SD < .75; Extremity = SD > 1.25

Appendix D-7. Indices of Centrality/Extremity for Listening, Internal v. External

		$r_{exp,res}$		$r_{meas,res}$	
Judge	Raw Score Standard Deviation	INTERNAL	EXTERNAL	INTERNAL	EXTERNAL
J01	11.0	0.15	-0.64*	0.16	-0.62*
J02	10.0	0.09	-0.61*	0.09	-0.59*
J03	8.1	-0.08	-0.70*	-0.07	-0.68*
J04	14.1	0.34*	-0.45*	0.33*	-0.43*
J05	8.9	-0.05	-0.72*	-0.05	-0.70*
J06	8.6	-0.19	-0.78*	-0.18	-0.77*
J07	10.0	-0.37*	-0.79*	-0.37*	-0.79*
J08	8.7	0.11	-0.69*	0.11	-0.68*
J09	11.1	0.13	-0.65*	0.13	-0.65*
J10	13.7	-0.03	-0.55*	-0.03	-0.55*
J11	11.4	-0.45*	-0.78*	-0.44*	-0.76*
J12	10.2	0.05	-0.65*	0.05	-0.65*
J13	8.6	-0.19	-0.76*	-0.19	-0.77*
J14	12.2	0.15	-0.50*	0.13	-0.51*
J15	12.2	0.22	-0.53*	0.21	-0.53*
J16	12.1	0.32*	-0.52*	0.30	-0.54*
J17	9.9	0.13	-0.74*	0.13	-0.72*
J18	9.4	0.06	-0.73*	0.06	-0.72*
*Centrality		2	18	2	18
No Effect		14	0	15	0
*Extremity		2	0	1	0

*$p < .05$

Appendix D-8. Judge Fit Indices of Centrality/Extremity for Listening, Internal v. External

		InFit Mean Square		SD Residuals	
Judges	Raw Score Standard Deviation	INT.	EXT.	INT.	EXT.
J01	11.0	0.29*	1.06	0.73*	1.21
J02	10.0	0.30*	0.61	0.73*	0.92
J03	8.1	0.28*	0.89	0.66*	1.04
J04	14.1	0.51*	1.14	0.97	1.27*
J05	8.9	0.31*	1.20	0.75	1.30*
J06	8.6	0.22*	1.02	0.65*	1.23
J07	10.0	0.47*	1.57*	0.98	1.57*
J08	8.7	0.25*	1.12	0.66*	1.22
J09	11.1	0.28*	0.98	0.74*	1.22
J10	13.7	0.57*	0.99	1.11	1.27*
J11	11.4	0.89	2.37*	1.32*	1.89*
J12	10.2	0.28*	0.71	0.73*	1.03
J13	8.6	0.23*	0.69	0.67*	1.00
J14	12.2	0.50*	1.01	0.93	1.17
J15	12.2	0.36*	0.76	0.83	1.05
J16	12.1	0.34*	0.87	0.77	1.09
J17	9.9	0.19*	1.25	0.59*	1.35*
J18	9.4	0.20*	1.06	0.58*	1.18
*Overfit/ Centrality		17	0	10	0
Acceptable		1	16	7	12
*Underfit/ Extremity		0	2	1	6

*For Infit Mean Square: Overfit = fit < 0.60; underfit = fit > 1.4.
For SD Residuals: Centrality = SD < .75; Extremity = SD > 1.25.