DOCUMENT RESUME

ED 333 033 TM 016 536

AUTHOR Myford, Carol M.

TITLE Assessment of Acting Ability.

PUB DATE Apr 91

NOTE 27p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991). For a related document, see TM 016 535.

PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS *Ability; *Acting; *Aesthetic Values; Art Criticism; Comparative Analysis; Drama; Evaluation Criteria; Evaluation Methods; *Evaluators; High Schools; High School Students; Interrater Reliability; Matched Groups; Performance Factors; Secondary School Teachers; *Student Evaluation; *Value Judgment; Videotape Recordings

IDENTIFIERS Experts; *Performance Based Evaluation

ABSTRACT

The aesthetic judgments of experts (casting directors and high school drama teachers), theater buffs, and novices were compared as they rated the videotaped performances of high school students performing Shakespearean monologues. Focus was on going beyond the determination of between-judge agreement to determine whether there were objective criteria that could differentiate the group ratings of the students. Three questions were posed: (1) whether the item calibrations for the experts, theatre buffs, and novices were significantly different; (2) whether the contestant ratings differed; and (3) whether the groups differed in the harshness with which they rated abilities. The judge sample (N=27) included nine experts, nine theatre buffs, and nine novices, with each expert being matched with a theatre buff and novice of the same sex and approximately the same age and level of education. All of the judges viewed eight high school students' videotaped performances of 2-minute long monologues twice, rated the videotapes, and completed the 36-item Judging Acting Ability Inventory developed for this study. One month later, each judge viewed the same eight videotapes of the student performances twice, and again completed the rating and sorting tasks. With a few exceptions, theatre buffs and novices were as capable as experts in using the rating standards when they were explicit and in comprehensible language. Experts did rate some performances differently, showing evidence of multiple criteria for judging a performance. Experts also rated performers more harshly, suggesting the application of more professional standards. Implications for the study of expertise in aesthetic judgment are discussed. Eight tables and four graphs illustrate the study. (SLD)

Assessment of Acting Ability

Carol M. Myford
Educational Testing Service, Princeton, NJ

Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL, April, 1991

Assessment of Acting Ability

Does expertise in making judgments about acting ability exist? If so, what is the nature of expertise in performing this task? How do the aesthetic judgments of experts differ from those of novices? In the past, researchers have studied whether experts show stronger agreement in their judgments about works of art than novices do. Valentine (1962), Child (1968, 1972) and Winner (1982) have reviewed the literature on inter-judge reliability as a criterion for expertise. Some studies provide evidence of strong between-judge agreement in the ratings of experts (e.g., Burt, 1934; Child, 1962; Dewar, 1938; Einhorn & Koelb, 1982; Farnsworth, 1969), while other studies reveal a lack of agreement between experts' ratings (e.g., Frances & Voillaume, 1964; Getzels & Csikszentmihalyi, 1969; Gordon, 1956; Gordon, 1923; Skager, Schultz & Klein, 1966). Whether the ability to maintain between-judge agreement is a useful criterion for detecting expertise in making aesthetic judgment appears questionable, given these equivocal findings.

Researchers have shown little interest in trying to pinpoint other criteria beyond between-judge agreement that might differentiate the ratings of experts from those of novices. A few researchers have investigated experts' abilities to replicate their ratings on a second occasion (e.g., Bamossy, Johnston & Parsons, 1985; Beard, 1978; Dewar, 1938; Einhorn & Koelb, 1982; Farnsworth, 1969; Gordon, 1923; Skager, Schultz & Klein, 1966). With the exception of Beard, the researchers presented test-retest reliabilities for experts but not for novices. The researchers found that experts reproduced their ratings with a high degree of accuracy (i.e., correlations in the range of 0.7 to 0.9, depending upon the individual study). However, since researchers did not present test-retest reliabilities for novices, there was no basis for comparison to determine whether novices could reproduce their ratings with the same degree of accuracy. Beard (1978) did gather data to allow such a comparison. Experts and novices in his study repeated the rating task a week later. When Beard compared the two sets of rating data, he found that experts had higher test-retest reliabilities than novices. Perhaps the ability to reproduce one's ratings may hold some promise as a useful criterion for identifying expertise in aesthetic judgment.

Given Beard's findings, the question arises are there other criteria beyond between-judge agreement that might differentiate the aesthetic judgments of experts from those of novices? If so, what might those criteria be?

Statement of the Problem

The purpose of this study was to compare the aesthetic judgments of experts (i.e., casting directors and high school drama teachers), theater buffs, and novices as they rated high school students' videotaped performances of Shakespearean monologues. The judges repeated the rating task one month later for test-retest purposes. The study sought to determine whether there are objective criteria which could differentiate the three judge groups' ratings of the students (i.e., contestants). The goal of the study was to go beyond the investigation of between-judge agreement to search for other objective criteria that constitute "some necessary, if not sufficient, conditions for defining expertise within a given situation" (Einhorn, 1974, p. 562).

In the past, researchers have studied only a few criteria that they hypothesized should distinguish the ratings of experts from those of novices. The statistical tools available for analyzing rating data limited the kinds of criteria they could study. Recent advances in rating scale analysis methodology (Wright & Masters, 1982) now make it possible to gain an indepth understanding of many aspects of rating data that heretofore were not amenable to study. In particular, with the introduction of the FACETS rating scale analysis program (Linacre, 1989) researchers now have access to a powerful new statistical tool that can help them make sense of complex multi-faceted rating situations.

In the present study, there are four "facets" of the data that are of interest: (1) the rating items, (2) the contestants, (3) the judges, and (4) the rating occasions. Nine criteria derived from these four facets were posed as potential indicators of expertise. In this paper we will examine three of the criteria[1]. Each criterion was framed in the form of a question. For each criterion a conceptual explanation is included as well as a discussion of how the criterion may function as an indicator of expertise. The FACETS program produces a measure of each of the criteria. An explanation of each of these Rasch measures is included to show the direct linkage between the criteria posed and the statistical methodology employed.

Criterion 1: Are the item calibrations for experts, buffs, and novices significantly different? When the three judge groups use a set of rating items to judge actors' abilities, they may not share a common understanding of each item's meaning. Experts might define individual items differently than buffs and novices who have considerably less practice using such items. For example, novices might give high ratings on an item which they consider an "easy" item for high school students to master, while experts might give low ratings on the same item because from their experience they know that the item is a "hard" one for students to master. In this instance, the two groups do not share a common understanding of the item's meaning, and consequently

[1] The interested reader should consult Myford (1989) for an explanation of the other six criteria.

they use the item in different ways. The item's "difficulty" differs across groups.

When rating data are analyzed using the FACETS program, a measure of each item's difficulty called its "calibration" is computed from the judges' ratings on that item. The higher the calibration, the more difficult the item (i.e., the harder it is for a contestant to get a high rating on the item). The items on the rating instrument were calibrated separately for each judge group, and the three sets of calibrations were compared to determine whether there were items which the three groups used differently.

Criterion 2: Are the contestant measures for experts, buffs, and novices significantly different? When judges rate contestants, they will give some performances higher marks than others. The contestants can be ordered by ability from lowest to highest to describe a continuum of acting ability. Experts might rate contestants differently than buffs and novices. Their contestant ordering may differ from buffs' and novices' orderings. The groups may not define good acting in the same manner. What one group considers good acting another group might consider poor acting.

The FACETS program produces an estimate of each contestant's ability in logit units called a contestant "measure" which is computed from the judges' ratings of the contestant. The higher the contestant measure, the greater the contestant's ability. Contestant measures were computed separately for each group of judges, and the three sets of contestant measures were compared to determine whether there were performances which the groups rated differently.

Criterion 3: Do experts, buffs, and novices differ in the harshness with which they rate? When judges rate contestants, they may not all rate with the same degree of harshness or severity. While two judges may share a common understanding of the standards they employ, one may apply those standards with greater severity, giving contestants consistently lower ratings than the other. Perhaps experts as a group differ from buffs and novices in the level of harshness they exhibit when rating contestants.

A FACETS analysis provides a measure of judge harshness for each judge called a judge "calibration." The higher the judge calibration, the more harsh the judge. Mean judge calibrations for experts, buffs, and novices were compared to determine whether one judge group rated significantly more harshly than another.

Method

Subjects

The judge sample (*N* = 27) was composed of nine experts, nine theater buffs, and nine novices. A matched subjects design was employed. Since the subjects were not randomly selected, matching was used to control for the effects of age, sex, and educational level across the three groups. Each expert was matched with a buff and novice of the same sex and approximately the same age and level of education.

Experts in this study were casting directors and high school drama teachers practiced in their craft who had logged many hours in evaluating actors' abilities. Each had formal training in drama and was fluent in the language of the discipline. The experts were very familiar with the criteria used in judging acting ability and made such judgments routinely as part of their job assignments. They had experience working with actors of various ages and abilities including teenage actors.

Theater buffs who participated in the study were not formally trained in the discipline but attended professional theater regularly, read reviews, enjoyed talking about drama, and had some knowledge of the kinds of criteria used in evaluating acting. While they may have spent time discussing with others the merits and shortcomings of actors they had seen, they had neither the breadth nor depth of experience in critically analyzing performances that the experts had. Furthermore, while all the buffs attended professional productions, they infrequently viewed high school productions. It was hypothesized that the buffs represented an intermediate stage in the development of expertise in judging acting ability.

Novices in this study were persons who attended the theater very infrequently, rarely read critics' reviews of theatrical performances, and had little training or experience in drama beyond high school. They lacked knowledge of the technical vocabulary used in talking about acting and had no formal experience judging actors' abilities.

Materials

Videotapes

The judges rated eight high school students' videotaped performances of monologues from Shakespearean tragedies and history plays. Each monologue lasted approximately two minutes. All contestants' videotapes conformed to certain standards in order to control for extraneous differences between them (e.g., no character costumes, makeup, or changes in lighting, etc.). All contestants were taped against a neutral backdrop using one fixed camera at a fixed angle with a fixed lens.

The eight monologues were copied on to four master tapes. All tapes contained the same monologues, but the order of the monologues differed across tapes to counterbalance the presentation of the monologues across judges.

Judging Acting Ability Inventory

The judges rated monologue performances using the investigator-designed Judging Acting Ability Inventory which consists of 36 items, each item describing a standard of good acting. Eleven items are designed to assess the actor's voice. Eleven items assess the actor's body, and fourteen items assess the actor's characterization. Judges determine whether the student performs well or poorly on each standard and then decide how well or how poorly. All items use a common six-point rating scale with the points defined as "very poorly," "moderately poorly," "slightly poorly," "slightly well," "moderately well," and "very well." Judges circle their response to each item.

Procedure

Each judge met individually with the investigator for an hour. The judges viewed the performances twice--once to become familiar with the actor and the monologue, and the second time to rate each performance. Two tapes were used to counterbalance the presentation of monologues across judges. The investigator stopped the videotape after presenting each monologue to allow the judge to fill out the Judging Acting Ability Inventory for the contestant. After rating the eight performances, the judge sorted them into categories and then ordered the performances within each category from best to worst.

Each judge returned for a second rating session one month later to gather data to examine the question of replicability. Again, each judge saw the eight performances twice: the first time to become re-acquainted with the monologues, and the second time to rate the performances. The tapes were counterbalanced in the second session as in the first. The judge then sorted the performances into categories. After completing the rating and sorting tasks, the judges filled out a short questionnaire describing their education and experience in drama.

Results

Experts' buffs', and novices' ratings differed in several ways. In the following discussion, the results obtained for the three criteria are presented.[2]

Criterion 1: Are the item calibrations for experts, buffs, and novices significantly different? An omnibus chi-square test for rating consistency (an analogue to Hedges & Olkin's (1985, p. 123) test for homogeneity of effect sizes)[3] was run to determine whether the three sets of item calibrations were significantly different. The chi-square test revealed that the item calibrations for experts, buffs, and novices are significantly different ($\chi_{70}{}^2 = 125.08$, $p < .005$). Pairwise tests were run to determine where the between-group differences lay. The results showed that the experts' and buffs' item calibrations are significantly different ($\chi_{35}{}^2 = 73.08$, $p < .001$), buffs' and novices' item calibrations are significantly different ($\chi_{35}{}^2 = 58.79$, $p < .01$), and experts' and novices' item calibrations are significantly different ($\chi_{35}{}^2 = 53.62$, $p < .025$).

[2] The interested reader is referred to Myford (1989) for a discussion of the results obtained for the other six criteria not covered in this paper.

[3] Chi-square tests for rating consistency were used rather than traditional analysis of variance methods to test for significant differences in the three groups' item calibrations. Each item calibration has a standard error associated with it, and the computation of the chi-square statistic takes into consideration each item's standard error. By contrast, analysis of variance techniques assume that the error variance for the items is distributed identically and independently over all the calibrations, not acknowledging that individual items may have different standard errors. Because the chi-square test for rating consistency makes use of more information about each item (i.e., both the difficulty measure and the standard error for the measure), this methodology was selected over traditional analysis of variance techniques. The formula used to compute the chi-square statistic is presented in Myford (1989).

Which particular item calibrations are different across the three groups? A chi-square test for rating consistency was run for each individual item to pinpoint those particular items. The results showed that the three groups differed in their use of four items: Item 2 (Understands the meaning of the lines), Item 4 (Produces an unstrained tone), Item 6 (Speaks without regional dialects or affectations), and Item 22 (Integrates movement and text; actions suit words). Figures 1, 2, and 3 identify these outlier items. The experts' and buffs' calibrations for items 4, 6, and 22 were significantly different, while novices' and buffs' calibrations for items 2, 4, and 6 were significantly different. Experts' and novices' calibrations for item 2 were significantly different. The remaining 32 items have caiibrations that are not significantly different across the three judge groups.

Insert Figures 1, 2, and 3 about here

Criterion 2: Are the contestant measures for experts, buffs, and novices significantly different? Omnibus tests of rating consistency were run to determine whether the contestant measures vary significantly across groups at Time 1 and at Time 2. The contestant measures for the three judge groups were significantly different both at Time 1 ($\chi_{16}{}^2 = 593.12$, $p < .001$) and at Time 2 ($\chi_{16}{}^2 = 599.46$, $p < .001$).

Pairwise tests for rating consistency were run to determine where the between-group differences lay. The results displayed in Table 1 show that each groups' contestant measures were significantly different from the other two groups' contestant measures for both rating occasions. The largest difference was between experts' and novices' contestant measures, while the smallest difference was between experts' and buffs' measures. Buffs' measures of contestant ability were more like the experts' measures than the novices' measures at both Time 1 and Time 2.

Insert Table 1 about here

Which contestants did the groups rate differently? A chi-square test for rating consistency was run for each individual contestant to pinpoint those particular contestants whom the three groups viewed differently. Tables 2 and 3 present the results of those analyses. The chi-square values have been converted into *z* scores by taking the square root of each chi-square value. (The same information is presented in Figure 4 but in a pictorial format that more clearly displays the continuum of contestant ability. In Figure 4 each contestant measure is bracketed by its standard error.) The three groups differed in the estimations of various contestants' abilities as shown in Tables 2 and 3. For Time 1, 21 of the 24 between-group comparisons were significantly different; and at Time 2, 16 of the 24 comparisons were significantly different.

Insert Tables 2 and 3 about here

How did the groups' contestant measures differ? Did they order contestants by ability differently? To the contrary, Figure 4 shows that the contestant orderings for the three groups were similar. Each group's ordering shows a progression from Mercutio and Paulina at the lower end of the acting ability continuum to Caliban, Ophelia, and Mark Antony at the upper end of the continuum. Only in the case of the Lady Anne portrayal was there a decided difference of opinion about the placement of this performance in comparison to the others. With the exception of the Lady Anne performance, then, the groups seem to share a common definition of what constitutes "good" and "poor" acting.

Where the groups seem to differ is in their judgments of just how good or how poor a performance is. This is particularly noticeable in the cases of the Lady Anne, Mark Antony, and Ophelia performances. For these three contestants the novices' ratings were markedly higher than the buffs' and experts' ratings.

Insert Figure 4 about here

Criterion 3: Do experts, buffs, and novices differ in the harshness with which they rate? The judge calibrations of experts, buffs, and novices were compared to determine whether one judge group rated significantly more harshly than another. Table 4 displays the means and adjusted standard deviations of judge calibrations for experts (casting directors and drama teachers), buffs, and novices.

Insert Table 4 about here

Two one-way analyses of variance were run using the judge calibrations from Time 1 and Time 2. The results of the analyses are summarized in Tables 5 and 6. The means of the judge calibration distribution for experts (casting directors and drama teachers) are significantly different from the means of buffs and novices ($F = 5.30$, $df = 1/23$, $p = .03$) at Time 2 and approach significance at Time 1 ($F = 2.58$, $df = 1/23$, $p = .12$). Experts as a group rated contestants significantly more harshly than buffs and novices did.

Insert Tables 5 and 6 about here

Which judges rated more harshly than others? Tables 7 and 8 show the calibrations for each judge at Time 1 and at Time 2. The judges did not all rate contestants with the same degree of severity. Thus, the same contestant might receive significantly different ratings depending upor which judge rated that contestant. The judges were not interchangeable. In each judge group there were some judges who rated more harshly and others who rated more leniently.

Insert Tables 7 and 8 about here

Discussion

What do the results of this study tell us about the nature of expertise in making aesthetic judgments? In this study the three judge groups employed the 36-item Judging Acting Ability Inventory to rate high school students' performances on two different occasions one month apart. Each item described a standard of good acting. There were four items on the inventory which the groups used differently, but the groups shared a common understanding of what each of the other 32 items meant and employed each of those items consistently when judging performances. The investigator had hypothesized that buffs and novices, lacking experience with such specific standards, would not understand the rating items and would use them differently than experts would. However, this was not the case. There were only a few items that did not seem to convey the same meaning across the three groups.

Experts gave contestants lower ratings on Item 2 (Understands the meaning of the lines) and Item 22 (Integrates movement and text; actions suit words) than novices and buffs did. Why might experts have used these particular items differently than novices and buffs? Novices and buffs have had less prior exposure to the monologues and therefore would lack sufficient knowledge of the text to be able to determine whether actors understood the lines or whether they integrated movements and text. By contrast, experts, having an intimate knowledge of the monologue and the function of the monologue within the larger context of the play, would have developed expectations of how an actor ought to convey understanding of the lines and what kinds of actions are most appropriate for a given text. The groups used two other items differently as well: Item 4 (Produces an unstrained tone), and Item 6 (Speaks without regional dialects or affectations). These two items employ specialized terms which would be familiar to experts but might be unfamiliar to buffs and novices. Experts have mastered the technical vocabulary of drama, which Perry (1984) characterizes as a highly specialized use of language involving "the personal articulation of things very difficult to say" (p. 30). In short, they have learned how to think and talk about drama. Buffs and novices lack familiarity with the technical vocabulary. If buffs and novices did not understand the term "unstrained" or were unaccustomed to listening for affectations of speech, then they would have difficulty employing these two items.

With a few exceptions, then, buffs and novices were as capable as experts of using the rating standards when those standards were made explicit and were couched in language that buffs and novices could understand.

Where the groups differed was in how harshly they employed the standards and how they used the standards to rate certain contestants' performances. These differences functioned as indicators of expertise in the performance of the rating task.

Experts rated some performances differently than buffs and novices did. While expert, buff, and novice judges differed little in their ordering of contestants by ability, they did differ in their judgments of just how much better (or worse) some contestants' performances were than others. There were three actors in particular that the three groups judged differently. All three actors chose monologues which have strong emotional appeal. The character in each monologue is in mourning. Novices gave these performances much higher ratings than buffs and experts did. Novices may have been overly impressed with the actors' abilities to show intense emotion and may not have been aware of technical shortcomings of the performance. Experts, having more familiarity with the monologues and the plays from which they were taken, could point out those shortcomings such as adding words, mixing up lines, or engaging in repetitive gesturing. They were less likely to have been "taken in" by the strong emotions. Rather, they would weigh the appropriateness and variety of emotions displayed, judging whether the emotions suited the words, not simply whether the actor could show emotion. Experts displayed evidence of decentering their perceptions, considering multiple criteria for judging a performance. By contrast, novices' judgments showed evidence of centering or focusing on a limited number of criteria when rating contestants. These findings provide support for the view that the expert sees more in a performance, knows what to focus his/her attention upon, and can look at a performance from a number of different angles.

Experts rated contestants significantly more harshly than buffs and novices, giving performances lower ratings than buffs and novices. Why would experts rate performances more critically than buffs and novices? The most obvious explanation is that experts are judging performances against "professional" standards. Experts are accustomed to viewing professional productions; so when they assess a high school student's performance, their frame of reference is the professional actor performing that monologue. Consequently, they would expect more than buffs and novices would, since buffs and novices have seen far fewer professional productions and lack a strong sense of what constitutes "good" acting. While this reasoning seems plausible, there is an alternative explanation that could be raised.

Perhaps the critical factor is not that the expert judges against "professional" standards, but rather that the expert more fully understands the capabilities of actors at various levels of ability. For any given ability level the expert has internalized a continuum that describes the highest (and lowest) levels of attainment that one can expect from an actor performing at that level of ability. Experts have worked with actors of various levels of ability and recognize that not all of them are capable of attaining their "vision of the ideal" (Altschuler & Janaro, 1967), so they adjust their expectations of each actor's capabilities taking into consideration factors such as the amount of training the actor has had, the actor's age, etc. Drama critic John Simon (as quoted in Searle, 1974) describes the importance of developing a "sliding scale" of excellence:

> A critical standard has to be both uniform and subdivisible. That is to say, in a sense you have a solid ideal of what you think is excellence. But, in another sense, you have a sliding scale and adjust it to the type of thing you are seeing . . . you sort of automatically evolve a sense of what might be the best that such a group could do, in your opinion, and then you judge according to that. (p. 11)

In Simon's view, the expert does not use different standards to judge persons of varying levels of ability. Rather, for each standard the expert can define low, medium, and high performance levels that reflect the expert's knowledge of what is the most and what is the least one can expect of an actor of a certain level of ability. Experts have a vast memory store of teenage actors' past performances. They have developed a finely tuned "yardstick" by which to assess the teenage actors' performances and are adept at placing individual performances along a high-school relevant ability continuum.

By contrast, buffs and novices have had little opportunity to develop such a memory store. Their frame of reference is limited. They have seen few high school performances and lack knowledge of the range of acting abilities high school students possess. Consequently, they do not have a realistic sense of just how much students at this age are capable of accomplishing.

This study makes several unique contributions to the literature on the nature of expertise in aesthetic judgment. First, the study moves us beyond the narrow focus of past research on inter-judge agreement as a criterion for expertise. In this study three other criteria were identified which proved useful in differentiating the ratings of experts from those of other judge groups. Second, the study investigated expertise in making judgments about acting ability. The vast majority of prior research on this topic has been in the visual arts, not the performing arts. This study extends the scope of research to a different arts domain. Third, past research has focused on comparing groups at both ends of the continuum--experts and novices. In the present study an intermediate group (i.e., theater buffs) was included which made it possible to study the transition from novice to expert. Finally, the present study employed data analysis techniques unlike those used in other studies of expertise in aesthetic judgment. Modeling the problem as a multi-faceted situation provided the means for investigating each of the facets independent of the other facets, making objective measurement possible.

References

Altschuler, T., & Janaro, R. P. (1967). *Responses to drama: An introduction to plays and movies*. New York: Houghton Mifflin.

Bamossy, G., Johnston, M., & Parsons, M. (1985). The assessment of aesthetic judgment ability. *Empirical Studies of the Arts, 3*(1), 63-79.

Beard, A. D. (1978). The quantification of visual aesthetic judgment using verbal and nonverbal scales (Doctoral dissertation, University of Chicago). *Dissertation Abstracts International, 39*, 3789A.

Burt, C. (1934). The psychology of art. In *How the Mind Works* (pp. 267-310). New York: D. Appleton-Century Co.

Child, I. L. (1962). Personal preferences as an expression of aesthetic sensitivity. *Journal of Personality, 30*, 496-512.

Child, I. L. (1968). Esthetics. In G. Lindzey & E. Aronson (Eds.), *The Handbook of Social Psychology: Vol. 3. The Individual in a Social Context* (2nd ed.) (pp. 853-916). Reading, MA: Addison-Wesley Publishing Co.

Child, I. L. (1972). Esthetics. In P. H. Mussen & M. R. Rosenzweig (Eds.), *Annual Review of Psychology: Vol. 23* (pp. 669-694). Palo Alto, CA: Annual Reviews, Inc.

Dewar, H. (1938). A comparison of tests of artistic appreciation. *British Journal of Educational Psychology, 8*, 149-158.

Einhorn, H. L. (1974). Expert judgment: Some necessary conditions and an example. *Journal of Applied Psychology, 59*(5), 562-571.

Einhorn, H. L., & Koelb, C. (1982). A psychometric study of literary-critical judgment. *Modern Language Studies, 12*(3), 59-82.

Farnsworth, P. R. (1969). Nature of musical taste. In *The social psychology of music* (2nd ed.) (pp. 97-133). Ames, IA: The Iowa State University Press.

Frances, R., & Voillaume, H. (1964). Une composante du jugement pictural: La fidelite de la representation. *Psychologie Francaise, 9*, 241-256.

Getzels, J. W., & Csikszentmihalyi, M. (1969). Aesthe'c opinion: An empirical study. *Public Opinion Quarterly, 33*(1), 34-45.

Gordon, D. A. (1956). Individual differences in the evaluation of art and the nature of art standards. *Journal of Educational Research, 50*(1), 17-30.

Gordon, K. A. (1923). A study of esthetic judgments. *Journal of Experimental Psychology, 6*(1), 36-43.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* New York: Academic Press.

Linacre, J. M. (1989). *Many-faceted Rasch measurement.* (Doctoral dissertation, University of Chicago). *Dissertation Abstracts International, 50,* 2029A.

Myford, C. M. (1989). *The nature of expertise in aesthetic judgment: Beyond inter-judge agreement.* (Doctoral dissertation, University of Chicago). *Dissertation Abstracts International, 50,* 3562A.

Perry, L. R. (1984). The arts judgment and language. *Journal of Aesthetic Education, 18*(1), 21-33.

Searle, J. (1974). Four drama critics. *Drama Review, 18*(3), 5-23.

Skager, R. W., Schultz, C. B., & Klein, S. P. (1966). Points of view about preference as tools in the analysis of creative products. *Perceptual and Motor Skills, 22*(1), 83-94.

Valentine, C. W. (1962). *The experimental psychology of beauty.* London: Metheun and Co., Ltd.

Winner, E. (1982). *Invented worlds: The psychology of the arts.* Cambridge, MA: Harvard University Press.

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

TABLE 1

PAIRWISE TESTS FOR RATING CONSISTENCY TO INVESTIGATE BETWEEN-GROUP DIFFERENCES IN CONTESTANT MEASURES

Groups	*Time 1* χ^2_8	*Time 2* χ^2_8
Expert vs. Buff	157.71*	104.96*
Expert vs. Novice	461.63*	531.81*
Buff vs. Novice	313.29*	285.05*

*$p < .005$

TABLE 2

DIFFERENCES BETWEEN CONTESTANT MEASURES FOR EXPERTS, BUFFS, AND NOVICES–TIME 1

Contestant	Expert Calibration	Buff Calibration	Novice Calibration	Exp/Buff z	Buff/Nov z	Exp/Nov z
Mercutio	-0.80	-0.22	-0.59	-8.20**	5.23**	-2.97**
Ophelia	0.48	0.65	0.93	-2.40*	-3.59**	-5.76**
Mark Antony	0.58	0.68	1.78	-1.41	-11.66**	-12.72**
Juliet	0.04	0.18	-0.49	-1.98*	9.48**	7.50**
Lady Anne	-0.24	0.23	0.70	-6.65**	6.02**	-12.04**
Caliban	0.60	0.99	0.98	-4.99**	0.12	-4.87**
Iago	0.67	0.52	0.51	2.12*	0.13	2.05*
Paulina	-0.27	-0.43	-0.69	2.26*	3.33**	5.38**

*$p < .05$
**$p < .01$

TABLE 3

DIFFERENCES BETWEEN CONTESTANT MEASURES FOR EXPERTS, BUFFS, AND NOVICES–TIME 2

Contestant	Expert Calibration	Buff Calibration	Novice Calibration	Exp/Buff z	Buff/Nov z	Exp/Nov z
Mercutio	-0.70	-0.42	-0.49	-3.96**	0.90	-2.69**
Ophelia	0.67	1.21	1.72	-6.28**	-5.15**	-12.21**
Mark Antony	0.64	0.60	1.61	0.57	-11.74**	-11.28**
Juliet	0.17	0.18	-0.03	-0.14	2.69**	2.56*
Lady Anne	-0.24	0.10	0.92	-4.81**	-10.50**	-14.85**
Caliban	0.78	1.12	1.07	-4.35**	0.59	-3.71**
Iago	0.49	0.67	0.62	-2.55*	0.64	-1.66
Paulina	-0.44	-0.51	-0.61	0.99	1.28	2.18*

*$p < .05$
**$p < .01$

TABLE 4

MEANS AND STANDARD DEVIATIONS FOR THE DISTRIBUTIONS OF JUDGE CALIBRATIONS

	Time 1		*Time 2*	
Group	Mean Calibration	Adj SD	Mean Calibration	Adj SD
Experts	-0.13	0.22	-0.17	0.31
Casting Directors	-0.16	0.31	-0.31	0.45
Drama Teachers	-0.12	0.10	-0.09	0.13
Buffs	-0.32	0.27	-0.37	0.27
Novices	-0.39	0.42	-0.60	0.33

TABLE 5

ANALYSIS OF VARIANCE ON JUDGE CALIBRATIONS FOR CASTING DIRECTORS, DRAMA TEACHERS, BUFFS, AND NOVICES–TIME 1

Source	*ss*	*df*	*ms*	*F*	*p*
Judge Type					
Directors vs. Teachers	0.02	1	0.02	.17	.72
(Directors + Teachers) vs. (Novices + Buffs)	0.31	1	0.31	2.58	.12
Novices vs. Buffs	0.02	1	0.02	.17	.69
Error	2.70	23	0.12		

TABLE 6

ANALYSIS OF VARIANCE ON JUDGE CALIBRATIONS FOR CASTING DIRECTORS, DRAMA TEACHERS, BUFFS, AND NOVICES–TIME 2

Source	*ss*	*df*	*ms*	*F*	*p*
Judge Type					
Directors vs. Teachers	0.17	1	0.17	1.70	.21
(Directors + Teachers) vs. (Novices + Buffs)	0.53	1	0.53	5.30	.03
Novices vs. Buffs	0.25	1	0.25	2.50	.14
Error	2.35	23	0.10		

TABLE 7

A COMPARISON OF JUDGE CALIBRATIONS FOR EXPERTS, BUFFS, AND NOVICES—TIME 1

Experts	Calibration		Buffs	Calibration		Novices	Calibration	
	Logit	Error		Logit	Error		Logit	Error
						Novice #19	0.48	0.06
			Buff #16	0.18	0.05			
Director #3	0.17	0.05						
			Buff #14	0.09	0.05			
Teacher #6	0.07	0.05						
Director #1	-0.01	0.05						
Director #4	-0.02	0.05						
						Novice #22	-0.04	0.06
						Novice #23	-0.05	0.06
Teacher #5	-0.11	0.05						
Teacher #8	-0.11	0.05						
Teacher #7	-0.19	0.05						
			Buff #12	-0.27	0.05			
			Buff #18	-0.29	0.05			
						Novice #26	-0.30	0.06
Director #9	-0.39	0.05	Buff #10	-0.39	0.05			
			Buff #11	-0.43	0.05			
						Novice #24	-0.51	0.06
Director #2	-0.59	0.05	Buff #17	-0.59	0.06			
			Buff #13	-0.60	0.06			
			Buff #15	-0.60	0.06			
						Novice #27	-0.70	0.06
						Novice #20	-0.72	0.07
						Novice #21	-0.77	0.07
						Novice #25	-0.88	0.07

TABLE 8

A COMPARISON OF JUDGE CALIBRATIONS FOR EXPERTS, BUFFS, AND NOVICES–TIME 2

	Calibration			Calibration			Calibration	
Experts	Logit	Error	Buffs	Logit	Error	Novices	Logit	Error
Teacher #7	0.23	0.05						
Teacher #6	0.15	0.05						
Director #3	0.09	0.05	Buff #14	0.09	0.05			
						Novice #19	0.08	0.06
Director #1	0.05	0.05						
			Buff #16	0.02	0.05			
Director #4	-0.15	0.05						
			Buff #13	-0.22	0.05			
Teacher #5	-0.33	0.05	Buff #18	-0.33	0.05			
Director #9	-0.38	0.05						
			Buff #12	-0.41	0.05	Novice #20	-0.41	0.06
Teacher #8	-0.43	0.05						
						Novice #22	-0.45	0.06
			Buff #15	-0.47	0.06			
			Buff #11	-0.49	0.06			
						Novice #23	-0.54	0.06
						Novice #26	-0.58	0.06
						Novice #24	-0.64	0.06
			Buff #10	-0.71	0.06			
Director #2	-0.76	0.06						
			Buff #17	-0.77	0.06			
						Novice #25	-0.84	0.07
						Novice #21	-0.85	0.07
						Novice #27	-1.16	0.07

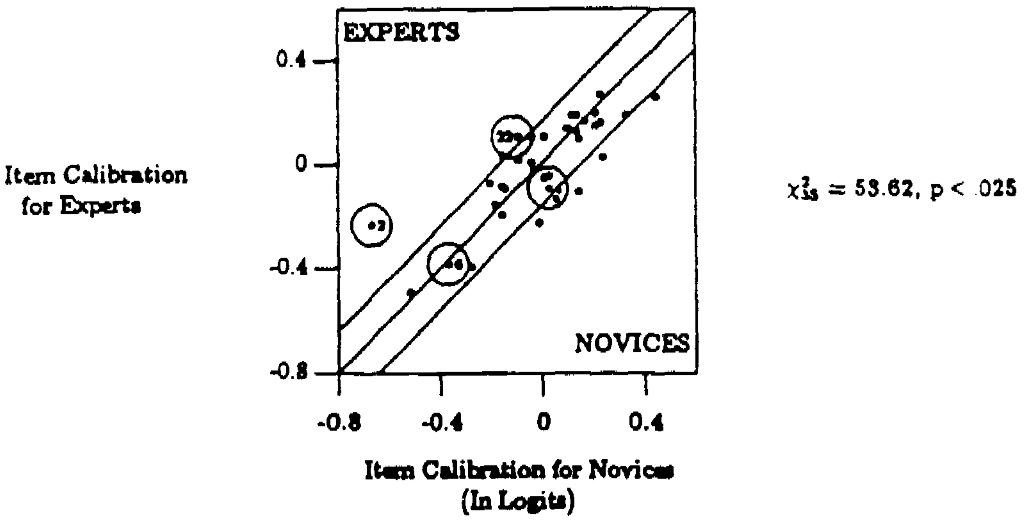

Figure 1. A comparison of item calibrations for novices and experts.

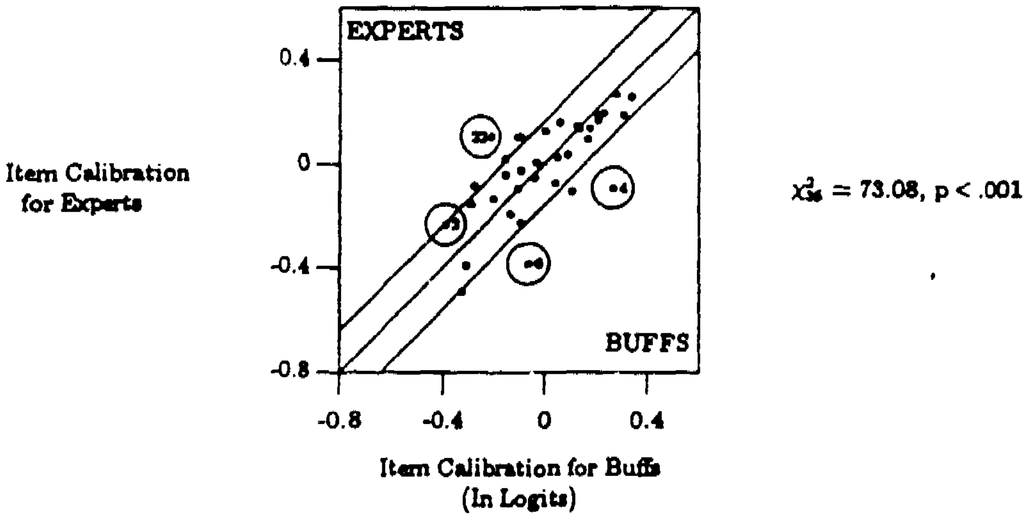

Figure 2 A comparison of item calibrations for buffs and experts.

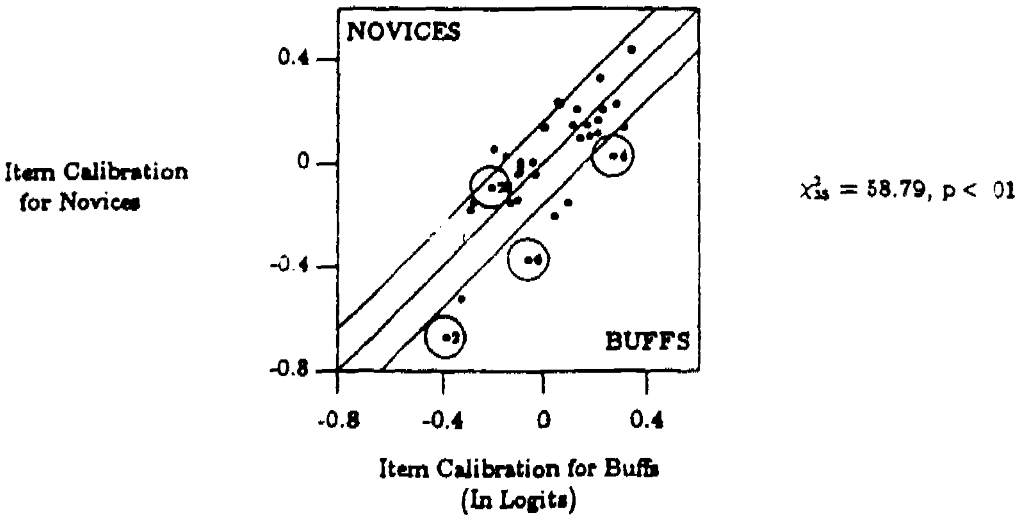

Figure 3 A comparison of item calibrations for buffs and novices.

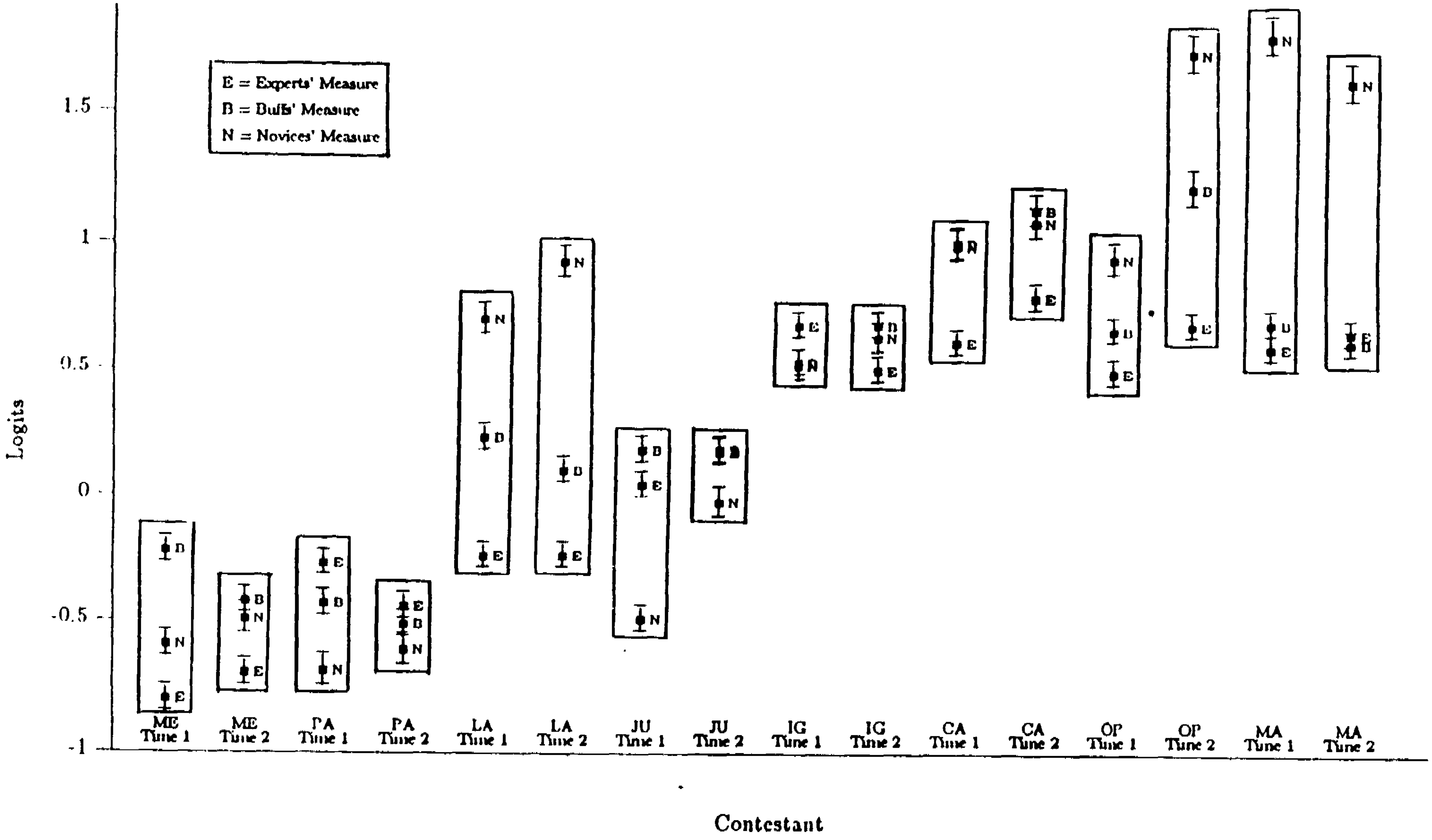

Figure 4 A comparison of novices', buffs', and exp ' measures of contestant ability at Time 1 and Time 2.