DOCUMENT RESUME

ED 228 324 TM 830 262

AUTHOR Ludlow, Larry H.; And Others
TITLE Measuring Change with the Rating Scale Model.
INSTITUTION Veterans Administration Hospital, Hines, IL. Rehabilitation Research and Development Lab.
PUB DATE 14 Apr 83
NOTE 27p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Attitude Change; *Attitude Measures; *Blindness; Longitudinal Studies; Program Evaluation; Rating Scales; Rehabilitation Centers; *Rehabilitation Programs; Test Interpretation; Veterans
IDENTIFIERS *Attitude Toward Blindness Questionnaire; Change Scores; *Rasch Model; Residuals (Statistics)

ABSTRACT

The Rehabilitation Research and Development Laboratory at the United States Veterans Administration Hines Hospital is engaged in a long-term evaluation of blind rehabilitation. One aspect of the evaluation project focuses on the measurement of attitudes toward blindness. Our aim is to measure changes in attitudes toward blindness from pre-training, to post-training, and then to a six month follow-up. The Attitude Toward Blindness Questionnaire consists of 39 statements scored as Strongly Agree:3, Agree:2, Disagree:1, Strongly Disagree:0. The Rating Scale Rasch model for ordered response categories was used to explain these data. The three periods were first calibrated as a single set, the residuals were analyzed, peculiar items were identified and removed. The analysis of change consisted of a calibration of items and persons at the separate time periods and the plotting of person attitude measures at each period. This paper demonstrates that: (1) detailed residual analyses can reveal critical measurement interaction processes, (2) measurement of change using the Rating Scale model is feasible, and (3) blind rehabilitation effects on attitudes can be studied quantitatively. (Author)

Measuring Change With the Rating Scale Model

by

Larry H. Ludlow
MESA Psychometric Laboratory,
University of Chicago, and
USVA Hines Hospital

Ross W. Lambert, Jr.
USVA Hines Hospital

Selwyn W. Becker
The Graduate School of Business,
University of Chicago, and
USVA Hines Hospital

Benjamin D. Wright
MESA Psychometric Laboratory,
University of Chicago, and
USVA Hines Hospital

This work was supported by the United States Veterans Administration:
Rehabilitation Research and Development Laboratory,
Hines, V.A. Hospital, Hines, Illinois.

Paper presented at the American Educational Research Association Annual Meeting. Montreal, Canada, April 14, 1983.
Paper printed in U.S.A.

Measuring Change with the Rating Scale Model

by
L.H. Ludlow, R.W. Lambert, Jr., S.W. Becker and B.D. Wright

ABSTRACT

The Rehabilitation Research and Development Laboratory at the USVA Hines Hospital is engaged in a long-term evaluation of blind-rehabilitation. One aspect of the evaluation project focuses on the measurement of attitudes toward blindness. Our aim is to measure changes in attitudes toward blindness from pre-training, to post-training, and then to a six month follow-up. The Attitude Toward Blindness Questionnaire consists of 39 statements scored as Strongly Agree:3, Agree:2, Disagree:1, Strongly Disagree:0. The Rating Scale Rasch model for ordered response categories was used to explain these data. The three periods were first calibrated as a single set, the residuals were analyzed, peculiar items were identified and removed. The analysis of change consisted of a calibration of items and persons at the separate time periods and the plotting of person attitude measures at each period. This paper demonstrates that: (a) detailed residual analyses can reveal critical measurement interaction processes, (b) measurement of change using the Rating Scale model is feasible, and (c) blind rehabilitation effects on attitudes can be studied quantitatively.

Purpose

The Rehabilitation Research and Development Laboratory at the USVA Hines Hospital is engaged in a long-term evaluation of blind rehabilitation. The Hines Blind Center accepts legally blind veterans on a voluntary basis in order to improve their mobility and overall quality of life. A wealth of qualitative data exist which attest to the efficacy of the Blind Center in the training of mobility skills, restoration of self-confidence, and improvement of attitude. The purpose of our evaluation is to measure changes in life which result from participation in the rehabilitation program.

The Center provides medical, mobility, social work and psychological services. Some patients require more assistance in one or another area. Overall, however, there is a common rehabilitation program in which all patients participate. One aspect of the evaluation project focuses on the measurement of attitudes toward blindness. The reason for studying attitudes is that the potential benefits of rehabilitation can be undermined by an unrealistic perception of one's limits and opportunities. Our aim is to measure changes in attitudes toward blindness from pre-training, to post-training, and then to a six month follow-up period. Three time points were chosen because it was hypothesized that if a program effect could be measured it would be strongest immediately after training and then perhaps decline somewhat over time but still have a noticeable effect at least six months after completion of training.

Instrument

The Attitude Toward Blindness Questionnaire consists of 39 statements scored as Strongly Agree:3, Agree:2, Disagree:1, Strongly Disagree:0. The instrument was pre-tested on 129 blind persons, rehabilitation workers, and "naive" persons with no familiarity with blindness. Two over-lapping forms of the instrument were constructed. A sub-sample of people took both forms, approximately two weeks apart. Some people received Form 1 first, others received Form 2 first. The Rating Scale Rasch model (Andrich, 1978; Wright & Masters, 1982) for ordered response categories was used to examine these data. The original questionnaire worked fairly well but a few group-by-item interactions were uncovered; some items were rejected, others were re-phrased. There was no form effect and, more important, there was no evidence of a learning effect for those who took both forms. These preliminary results led to the present form of the questionnaire and also the expectation that attitudes could be measured, and that attitudinal changes, if they occurred, could be attributed to the program and not to a test-retest learning effect.

The questionnaire consists of two sub-scales of related items. These scales address attitudes which reflect either a positive or negative consequence of blindness. This paper discusses the results for the Positive scale. Condensed versions of the items are provided in Table 1.

Data

In the present analysis 118 patients were measured at Time 1; 75 of them were remeasured at Time 2; 29 were remeasured at Time 3. Two males (A and C) and one female interviewer (B) collected the data; interviewers A and B at Time 1, interviewer C at Time 2, and interviewer B at Time 3. Interviewers received similar training and held meetings to discuss data collection anomolies. Time 1 interviews were conducted either by phone immediately prior to the patients' admission or immediately upon arrival at the hospital. The mean length of stay in the program was approximately three months. Time 2 data were collected by phone anytime from 2 to 6 weeks after release from the hospital. Time 3 data were collected by phone six months after the Time 2 interview.

Analysis

The Rating Scale model was used for this analysis. The model yields person and item location estimates and, in addition, category threshold estimates which indicate the difficulty of moving from one categorical response level to the next. The threshold estimates are computed across all items but the model expects the estimates to be valid for individual items as well. Threshold estimates can be disordered from one category to the next but from the design perspective ordered estimates are preferred. The analysis was accomplished with the computer program, CREDIT, developed by Masters, Wright and Ludlow

(1982).

The first analyses addressed the extent of item invariance across time; the variable must be the same at each time period. Two analytic approaches were possible. We could separately calibrate the data at each time period and then construct bivariate plots of item estimates. If the item points approximated a line with unit slope, then we could feel comfortable that items were operating in a similar manner at each time period. If the items exhibited dramatic shifts in relative location from one period to the next, however, then a discussion of attitude change might be meaningless.

A second approach is to calibrate the three time periods as a single set and then analyze the residuals by time period. If common item estimates are appropriate for the three time periods when they are calibrated together, then the vectors of residuals for each time period on each item should resemble one another in their distribution. If, however, an item is operating differently at one of the time periods, then the common item estimate will be inaccurate for the measures collected at that period. A negative residual pattern will occur when a common item estimate is too low for the people at that period. This is because expected scores on that item are too high and the observations are, therefore, less than expected. A positive residual pattern will occur when a common item estimate is too high. This is because expected scores are too low and observed scores are, therefore, higher than expected. Two advantages of the residual analysis are: a) the possibility of identifying people with unexpected

responses and, b) investigating potential sources of undesirable influence upon the measurement process. The residual analysis approach was used.

Table 1 contains the calibration results for the three time periods. The items are listed in their difficulty order. At the high end of the scale are items that are hardest to agree with. Only a person with a very positive attitude would say that being blind is an asset to marriage. At the low end are the easiest items. Only a person with a very negative attitude would say that a blind person could not raise a normal child. The threshold estimates are ordered: it is quite easy to move from strongly disagree to disagree, harder to move from disagree to agree and hardest to move from agree to strongly agree. The patient measures were symetrically distributed across the item locations. The mean measure was .81. Overall , the item and threshold orders are sensible and conform to the original intent of the scale. But, why then do some items misfit?

When a criterion was specified for flagging people with a fit statistic greater than ¶2.0¶, it was seen that for the data collected by interviewers A and B at Time 1 and interviewer B on Time 3, 18% of their sample had positive misfits and 13% had negative misfits. For the Time 2 data, taken by interviewer C, there were 3% positive misfits, and 23% negative misfits. Inspection showed that interviewer C tended to record mid-range responses while the other interviewers tended to record across the full range of responses. An analysis of residuals offers insight into how different interviewer recording

strategies can be revealed.

A tendency to avoid the extreme response categories is reflected in the category threshold estimates reported in Table 1. If interviewer C had an even greater tendency to avoid the extreme categories, then the expected scores for his patients could mean something different than the scores of other patients with the same observed total score. If the overall threshold estimates were considerably less extreme than those computed separately for his patients, then the residual pattern for his patients would tend to have a large standard deviation on most items because the probability of responding in an extreme category would be higher for his patients when combined in the overall analysis than if measured separately. This would lead to more extreme expected scores and, consequently, large residuals could occur. Actually, the overall threshold estimates are so large that his group of patients still appear to over fit the model because they have responded closer to expectation than would normally occur by chance. When the standard deviation of the residuals for each item were computed by interviewer and time period, it was found that in 87% of the comparisons, interviewer C's standard deviations were the smallest. His scoring pattern was consistent and conservative. We now had to consider the possibility and consequences of an interviewer and time period effect.

Figure 1 plots the residuals against the items' presentation order. There is a cluster of large negative residuals for some of the first items. The three misfitting items in Table 1 are the first,

second and seventh items in this figure. This configuration suggests some type of "start-up" effect. Figure 2 is a plot of residuals sorted by interviewer and time period for item I128, the first item presented to the patients. Five of the six largest negative residuals are due to interviewers A and B at Time 1. The combined standard deviation of residuals is largest for Time 1.

Discussion with interviewers revealed that most patients do not respond using the suggested category labels. Instead they respond "right", "false", "it depends", etc. The interviewers are required to make judgements about how to score those responses. After a few items they can usually pick up the patients' pattern and distinguish between middling and extreme responses. But each interviewer handles that situation in an idiosyncratic fashion. These ambiguous situations are the ones in which interviewers A and C were most interpretive in their response recording. Interviewer B tended to apply a more conservative criteria before an interpretive judgement was recorded as a "strongly" response. It was also noted that some patients actually do use the "strongly" labels but quickly revert to a more personalized mode of expression after a couple of items. Although some of this "interviewer start-up effect" diminished as interviewers gained more experience, it remains a systematic source of measurement error in the present analysis.

The second item in Figure 1, while also subject to an interviewer start-up effect, has a more likely explanation. The item states "a blind person can offer their spouse satisfactory sex". A plot of

residuals, similar to Figure 2, revealed that interviewer B recorded many disagree-type responses which, subsequently, produced large negative residuals. Men with otherwise positive attitudes were expressing a negative attitude on this item. Interviewer B is a woman. Is she eliciting the accurate life-state responses while the male interviewers are eliciting macho responses? Or is she eliciting inhibited responses while the men are eliciting the true condition? This single item offers no clue as to which situation is ocurring.

The third misfitting item from Table 1, N003, is plotted in Figure 3. Although Interviewer B stands out because of large negative residuals, an explanation other than interviewer gender is possible. When queried about why this item might misfit, all three interviewers agreed that the item content was ambiguous. The item states: "A blind person develops extra senses." To this statement most patients provide an agree-type response. Interviewer C at Time 2, however, doubted the reliability of these answers. After stating the item as written, he would re-phrase the item as "Do you mean you agree that new, previously non-existent senses emerge or do you mean that a blind person enhances existing senses?". Most patients would say that they meant existing senses were enhanced. Thus some patients with low positive attitudes gave high positive responses which produced positive residuals. Other patients with positive attitudes disagreed with this item and received large negative residuals. Large negative residuals would also be likely to occur more frequently after rehabilitation since the program emphasizes the enhancement of senses aspect. But if that is so why doesn't the Time 2 residual pattern

resemble the Time 3 pattern? One explanation is that some of the patients who were queried about the new versus enhanced senses distinction by interviewer C at Time 2 remembered the particular emphasis given that item and responded with a disagree-type response when they again encountered the item at Time 3.

Before an analysis of change began some remedy for these side-effects was necessary. The first item cannot be included in the present analysis because it is producing too much "start-up effect" error. The item regarding a sixth sense was removed because it elicited amibiguous responses from many patients. The item regarding sex, however, posed a thornier problem. If the effect of an interviewer's gender is restricted to that item, then the item could be removed. But what if the effect is more pervasive and interacts with other items? Figure 4 addresses the interviewer gender-by-item content issue.

Figure 4 contains a plot of pairs of residual means for each item for one pair of interviewers. If the data fit the model, then a random pattern should appear for pairs of items for any comparison of interviewers or time periods. A random pattern would be a meaningless cluster of points lying close to the origin. If there is no gender interaction effect then there should be no pattern when interviewer A means at Time 1 and interviewer B means at Time 1 are plotted against one another. In Figure 4 the horizontal axis represents interviewer A (the male), the vertical axis represents interviewer B (the female). This figure contains three points that stand out prominently from the

others. In Quadrant II the discrepancy in scoring for I128 has already been addressed in terms of the "start-up" effect. The relation between item I226 (about sex) and I220 (about marriage), however, is a new piece of information. Interviewer B, the woman, has elicited surprising negative responses from some patients on two items related to a common personal issue. A similar configuration resulted when her means were plotted against the other male at Time 2. When her means at Time 1 and Time 3 were plotted against each other items I226 and I220 were the only points in the third quadrant. In brief, the responses she and the male interviewers elicit on these two items are different from each other. In order to protect our change measures from this interviewer gender effect, these two items were removed from the estimation of patients' attitudes. Even though I220 did not misfit in the sense of producing a large fit statistic during calibration, we have found it subject to a systematic measurement error that reduces it's validity.

There was some doubt that the patterns in these residuals existed as a consequence of the deterministic factors claimed or whether the patterns could have occurred just as likely by chance. Simulation exercises were conducted using the item difficulties, category thresholds, and person measures obtained from the rating scale calibration. The response vectors generated from these estimates were identified by interviewer and time period and the residuals were sorted and analyzed in the same manner as the original data. In each case, the figures presented in this analysis were clearly divergent from the residual patterns resulting from data generated to fit the model, given the

original location estimates. Although the actual number of residuals to draw our attention to some aspect of the data is relatively small, confidence is placed in the interpretation of their deterministic origin. The simulations are available as a companion paper presented at this convention.

The analysis of change was based on the remaining 15 items. The items and patients were recalibrated across the time periods. This approach ensures that a raw score receives the same location estimate at each time period. There were 28 patients with Time 3 measures. A reasonable plot would show their attitude positions at each time period. However, there are missing data. Nineteen of these patients either missed Time 1 or Time 2 interviews. In most cases the missing data were systematic, e.g. patients would refuse to be interviewed. While this fact is useful, perhaps, for understanding an individual it cannot contribute full information about program overall effectiveness. We need to know where a patient started from before we can understand where he currently stands.

Figure 5 shows the measures for all patients who had Time 3 data. For these patients there as a Time 1 and Time 3 group, a Time 2 and Time 3 group, and a Time 1, Time 2 and Time 3 group. If we just look at the Time 1 and Time 3 group (broken line elipse) we do not know what happened at Time 2, which is important because their pattern suggests no change. If we just look at the Time 2 and Time 3 group it looks like their attitudes have declined after program participation but we do not know from where they started. The group present at all

three time points resembles the hypothesized pattern. That is, improvement from Time 1 to Time 2, slight decline from Time 2 to Time 3. If this were the true effect of the program, then we could infer the missing data pattern for the 1/3 and 2/3 groups.

Figure 6 contains the three measures for those 8 patients, who at the time of this analysis, had completed all interviews. The connecting lines illustrate the direction of change for each patient. Admittedly, the small sample of complete data does not present convincing evidence of program effectiveness but as more data are collected the patern should become sharper.

What does it mean to gain or lose in one's attitude? To the right of Figure 6 are two columns containing item locations determined by first adding the threshold estimate for the disagree/agree categories to the item estimates (agree column) and then adding the agree/strongly agree threshold estimate to the item locations (strongly agree column). These two columns indicate how affirmative a patient is expected to respond. For example, patient #2 is expected to agree with all items at Time 1. At Time 2 he has a 50% probability of strongly agreeing with N001. At Time 3 he is expected to strongly agree everything from N007 downwards.

Patient #8 is also noteworthy. At Time 2 he stated "he owed his life to Hines" and his self-confidence and motivation had been restored. By Time 3 he was divorced and had undergone severe insulin attacks while in a nursing home. He still credited the program but he

was no longer as upbeat about his present or future situation.

Conclusion

This analysis demonstrates that: (a) detailed residual analyses can reveal critical measurement interaction processes, (b) measurement of attitude change using the Rating Scale model is feasible, and (c) blind rehabilitation effects on attitudes can be studied quantitatively.

References

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

Masters, G.N. A Rasch model for partial credit scoring. Psychometrika, 1982, 47, 149-174.

Masters, G.N., Wright, B.D. & Ludlow, L.H. CREDIT: A Rasch program for ordered categories. Chicago: MESA Psychometric Laboratory, University of Chicago, 1982.

Wright, B.D. & Masters, G.N. Rating Scale Analysis. Chicago: MESA Press, 1982.

Measuring Change With the Rating Scale Model
HANDOUT

by

Larry H. Ludlow
MESA Psychometric Laboratory,
University of Chicago, and
USVA Hines Hospital

Ross W. Lambert, Jr.
USVA Hines Hospital

Selwyn W. Becker
The Graduate School of Business,
University of Chicago, and
USVA Hines Hospital

Benjamin D. Wright
MESA Psychometric Laboratory,
University of Chicago, and
USVA Hines Hospital

This work supported by the United States Veterans Administration: Rehabilitation Research and Development Center, Hines, V.A. Hospital, Hines, Illinois.

ABSTRACT

The Rehabilitation Research and Development Laboratory at the USVA Hines Hospital is engaged in a long-term evaluation of blind rehabilitation. One aspect of the evaluation project focuses on the measurement of attitudes toward blindness. Our aim is to measure changes in attitudes toward blindness from pre-training, to post-training, and then to a six month follow-up. The Attitude Toward Blindness Questionnaire consists of 39 statements scored as Strongly Agree:3, Agree:2, Disagree:1, Strongly Disagree:0. The Rating Scale Rasch model for ordered response categories was used to explain these data. The three time periods were first calibrated as a single set, the residuals were analyzed, peculiar items were identified and removed. The analysis of change consisted of a calibration of items and persons at the separate time periods and the plotting of person attitude measures at each period. This paper demonstrates that: (a) detailed residual analyses can reveal critical measurement interaction processes, (b) measurement of change using the Rating Scale model is feasible, and (c) blind rehabilitation effects on attitudes can be studied quantitatively.

Paper presented at the American Educational Research Association Annual Meeting, Montreal, Canada, April 14, 1983.
Paper printed in U.S.A.

Scoring Key:

Strongly agree=3 Agree=2 Disagree=1 Strongly disagree=0

Seq.	Item	Description	Value	SE	Fit
19	I220	asset to marriage.	2.35	.11	0.03
13	I133	do better telephone work	1.08	.11	-2.33
15	I118	are more honest than sighted	0.92	.11	0.99
16	I126	can endure boring tasks	.56	.11	-0.75
4	I219	don't superficially judge people	.55	.11	0.01
17	N009	closer to spouse than sighted	.53	.11	0.09
10	N005	blind workers complain less	.37	.11	-1.49
14	I119	blind worker distracted less	.32	.11	.63
12	I222	understand feelings better	.18	.11	-0.80
11	N007	especially loyal friend	-0.34	.12	-0.83
18	N012	can be good supervisors	-0.35	.12	-1.73
9	I120	unusually good negotiator	-0.60	.12	-4.69
5	I218	participate in group activities	-0.62	.12	0.21
7	N003	develops extra senses	-0.66	.12	3.36*
6	I131	superior piano tuners	-0.68	.12	-0.24
8	I122	sensitive social workers	-0.70	.12	-3.22
2	I226	offer spouse satisfactory sex	-0.75	.12	3.45*
1	I128	enjoy work as well as anyone	-1.08	.12	2.69*
3	N001	can raise a normal child	-1.09	.12	1.87

Threshold statistics:				
	Values	-2.53	-0.17	2.70
	Errors	.10	.04	.05

Table 1.--Calibration results for all items

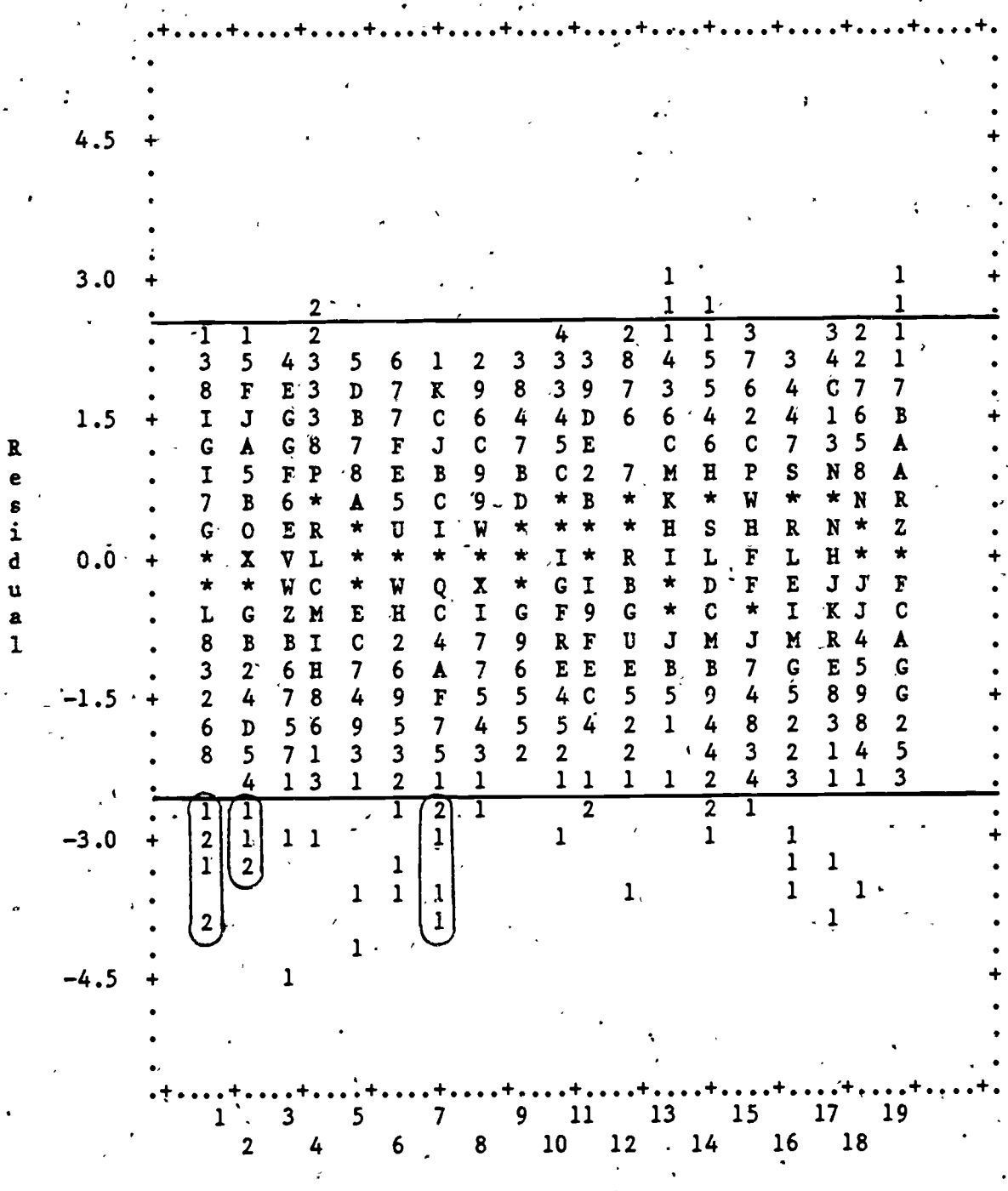


Figure 1.--Residuals plotted in item sequence order

Item I128

"enjoy work"

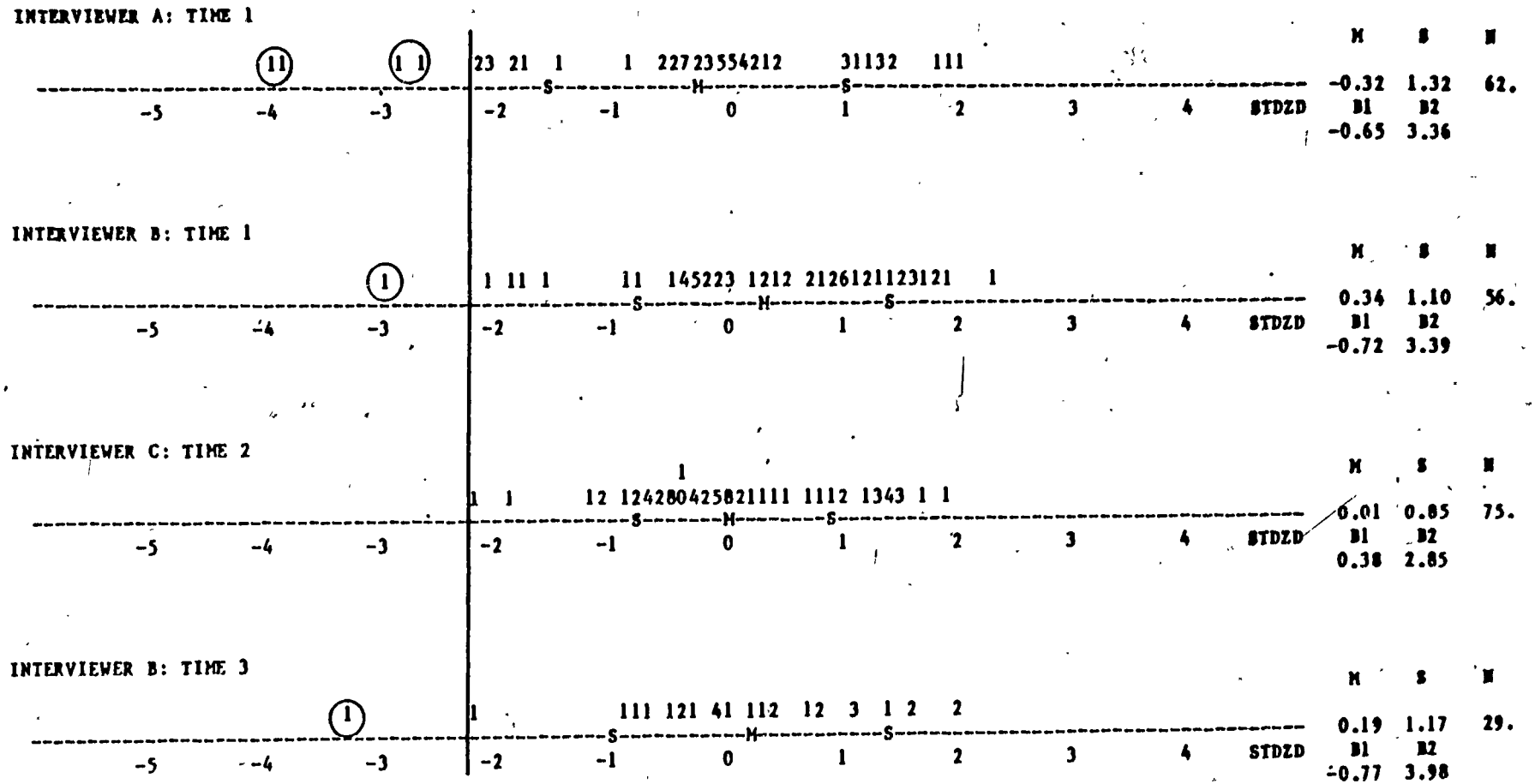


Figure 2.--Line plot of residuals by interviewer and time period: Item I128, calibration position= -1.08, fit= 2.69

Item N003

"extra senses"

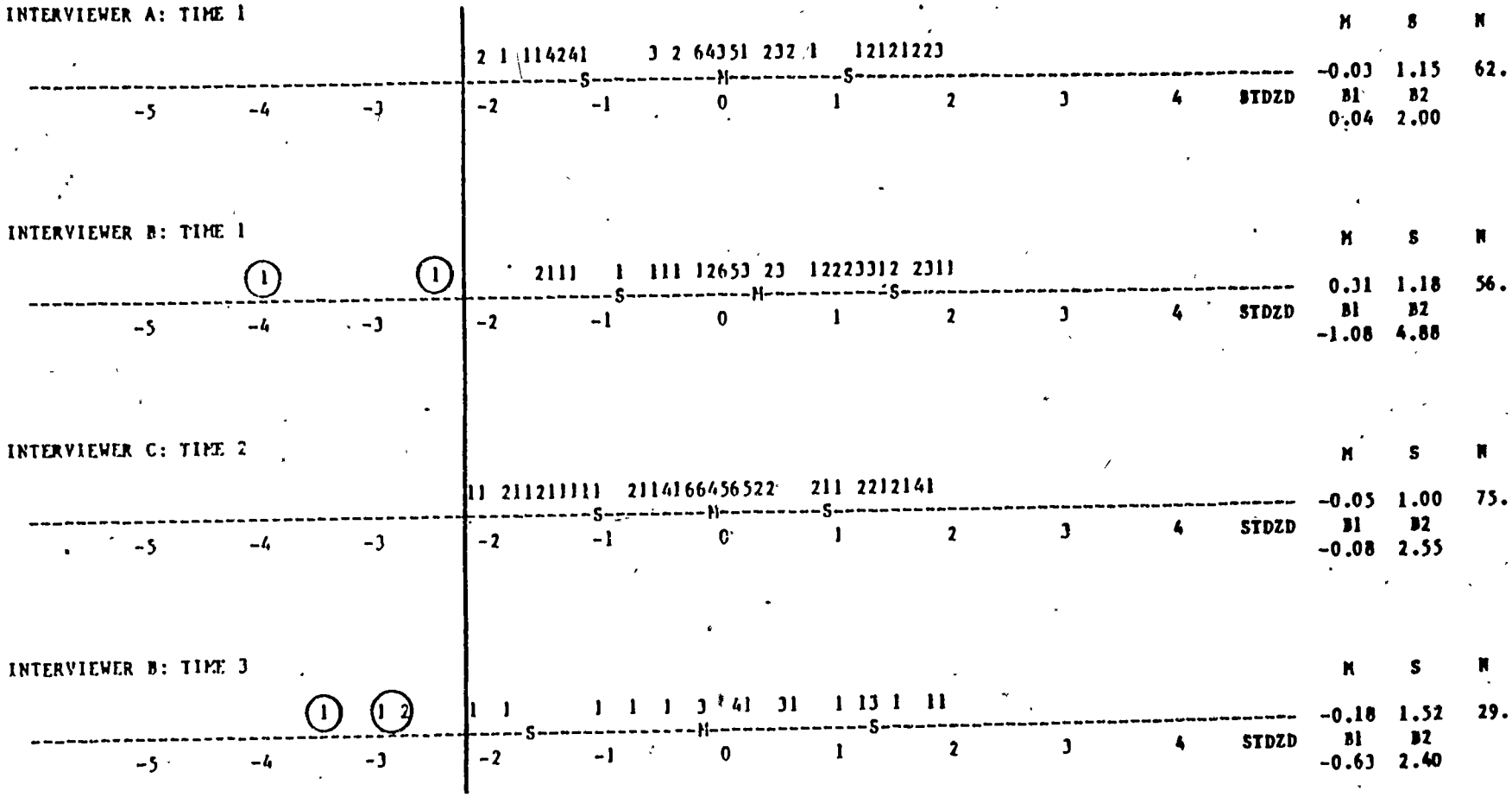


Figure 3.--Line plot of residuals by interviewer and time period: Item N003, calibration position= -0.66, fit= 3.36

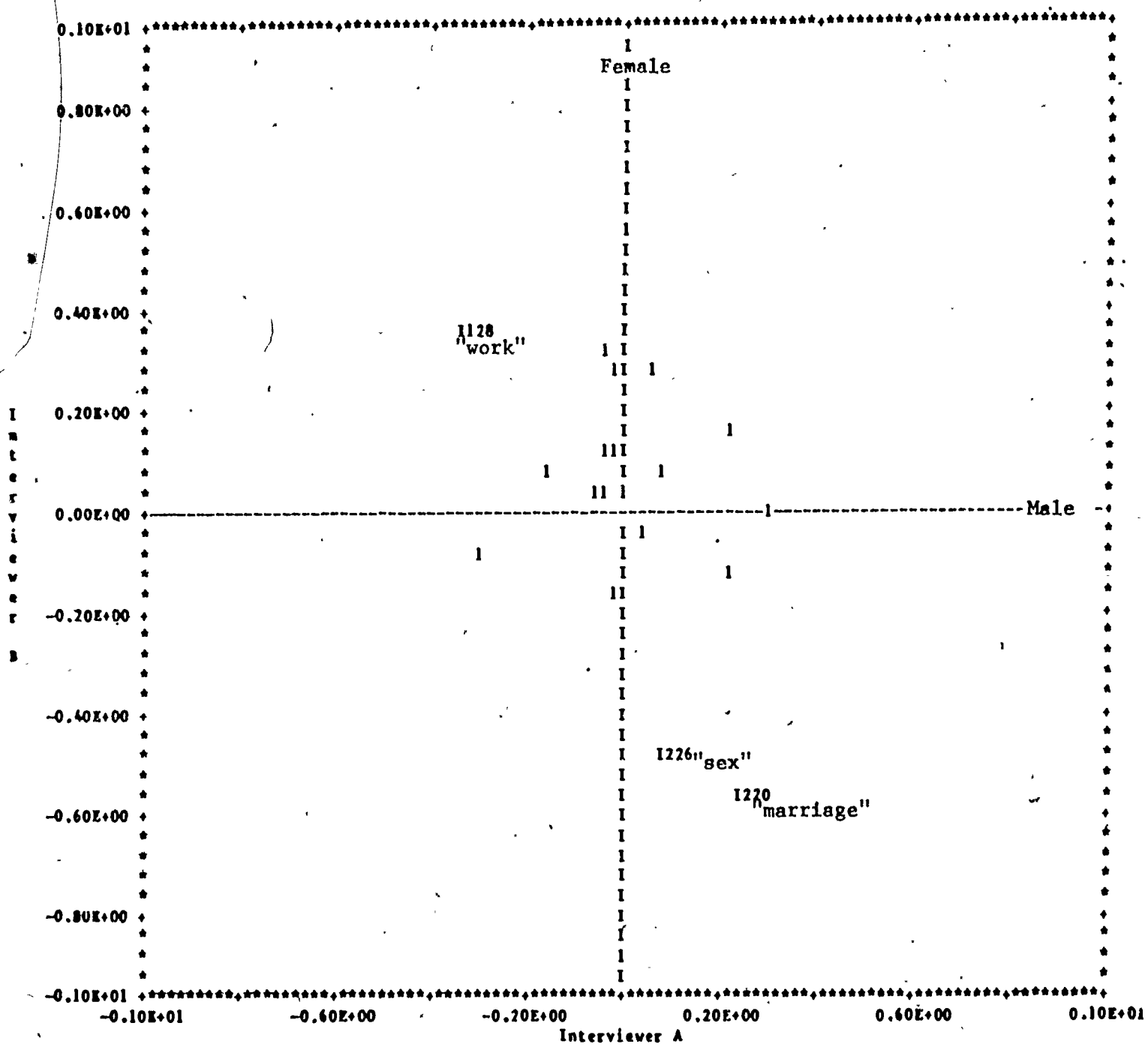
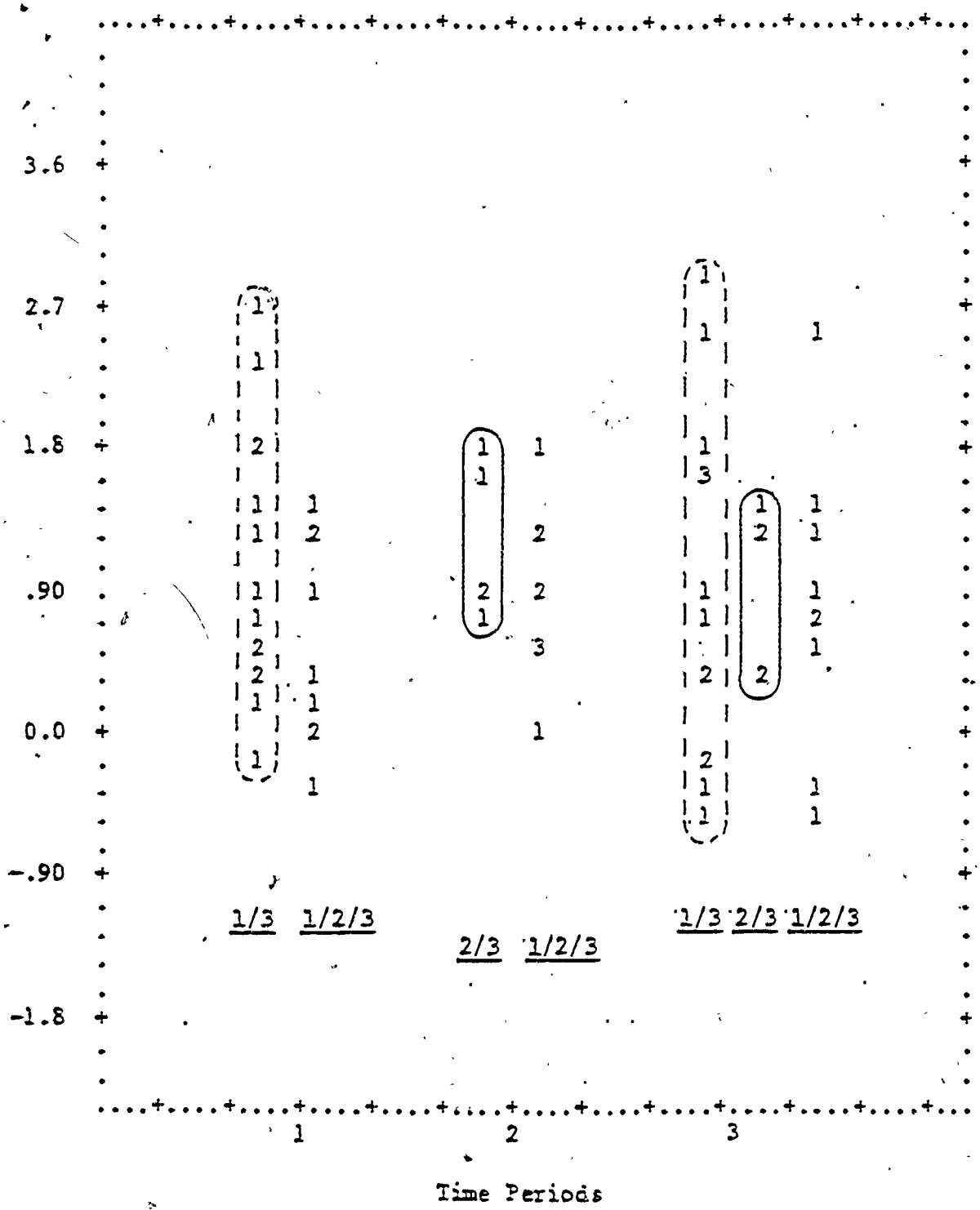


Figure 4.--Bivariate plot of mean residual on each item for Interviewer A and B at Time 1



Key

1/3	Persons measured at Time 1 and Time 3 only.
2/3	Persons measured at Time 2 and Time 3 only.
1/2/3	Persons measured at Time 1, Time 2 and Time 3.

Figure 5.--Measures at each time period: Sorted by number of time points actually collected.