

INTERRATER RELIABILITY IN AN IMPERFECT FIELD SETTING

DROR WALK

RACHEL FLEISHMAN

GAD MIZRAHI

The JDC-Brookdale Institute of Gerontology and Human Development

MIRIAM BAR-GIORA

The Service for the Aged of the Ministry of Labor and Social Affairs

In evaluating the methodology of a program, a conflict often arises between the pure nature of the theoretical framework and the dynamic features of the program as it is implemented. As a result, this type of evaluation poses a special challenge, unlike the evaluation of a survey or experimental study. Today I will present an analysis of an inter-rater reliability test, which was conducted on the questionnaires used in the surveillance of residential care institutions. We attempted to quantify the variables that exist in the natural setting in which we conducted the reliability test. This type of evaluation is especially important, because it can lead not only to an improvement in the tools, but also to a better understanding of the use of reliability tests in the field.

DESCRIPTION OF THE PROJECT

In the late 1980s, the Service for the Aged of Israel's Ministry of Labor and Social Affairs, and the JDC-Brookdale Institute, began implementing an experimental program.

The aim of this program was to improve the surveillance of residential care institutions for the semi-independent and frail elderly, in the hope that this would then improve the quality of institutional care. This program used the RAF Method of **R**egulation, **A**ssessment, **F**ollow-up and **C**ontinuous Improvement of Quality of Care. The RAF Method is based on quality assurance principles, as well as on the tracer approach to evaluating the quality of care.

During surveillance, information about the quality of care is gathered from several sources in an institution. These include interviews with residents and a review of patient records.

Surveillance covers many areas of care, like infrastructure and nursing care. While evaluating this experimental program, the reliability of the surveillance tools was also tested.

I will discuss the findings of the reliability test, using two units of analysis: The first is the standard unit used in reliability analyses – that is, the questions in the questionnaire. The second is a unit of analysis that reflects the imperfect nature of the field setting.

METHODS

The reliability measurement

Cohen's kappa coefficient is the most widely used measure of inter-rater reliability in the social sciences. However, it is difficult to apply, since it relies on the multiplicative law of probability to estimate the likelihood of agreement by chance, consequently “punishes” asymmetrical distribution of responses. Because of that, kappa can only be used when it is reasonable to assume *a priori* that raters will assign a given proportion of cases to a given category. In most applied research, however, there is no justification for such an assumption, and such an assumption can be especially misleading if there is a strong imbalance in the prevalence of different response categories.

Perreault and Leigh proposed an alternative to Cohen's kappa for the two-rater case. Their measure assumes that the agreement observed among raters is a function of a true (unknown) level of reliability. Consequently, the true level of reliability can be inferred from the level of agreement, because of the monotonic mapping between the two. We can calculate the resulting measure as follows:

$$I_r = \begin{cases} \{[A - (1/K)] [K/(K - 1)]\}^{1/2} & \text{if } A \geq 1/K \\ 0 & \text{if } A < 1/K \end{cases}$$

where A is the observed proportion of inter-rater agreement, and K is the number of categories into which the responses can be coded.

In this study, we use Perreault and Leigh's I_r as the reliability measure. Perreault and Leigh's I_r is “better behaved” than Cohen's kappa, because it always achieves a maximum of 1 when there is perfect inter-rater agreement. It is defined as zero when the number of agreements is less than or equal to what we would expect to occur by chance.

Procedure

We conducted the reliability test by having the surveillance questionnaires be administered by teams of two surveyors (a social worker and a nurse) on two separate occasions. The questionnaires contained 257 questions. Ninety percent of them had two valid response categories, and 10 percent of them had three valid response categories. On average, 15 days elapsed between the two inspections. Thirty-two long-term care institutions were inspected. They had a total of 2,213 residents, and constituted a representative sample of long-term care institutions in Israel.

The first time, the questionnaire was administered by several different surveillance teams, as part of the routine government inspection. The second time, the questionnaire was administered again by an outside team, whose members had experience working with elderly residents of institutions.

As I've said, designing an inter-rater reliability test for a natural setting is especially challenging, because of the difficulty of meeting methodological standards. Since this reliability test was conducted as part of routine inspection, and not as part of a controlled experiment, it was liable to deviate from the optimal model of a reliability test. For example,

it was possible that an institution's director or residents would refuse to be inspected twice within a short time. This forced us to schedule the two inspections at least ten days apart. Moreover, at all of the institutions, the second inspection was conducted by an outside team, and not by any of the regular surveillance teams that had conducted the first, "real" inspection.

Because this study was conducted in a natural setting, and therefore was not "pure", it was difficult to attribute variance between the two inspections only to the nature of the tool. Nevertheless, it gave us an opportunity to measure the impact on reliability of the variables that characterize a reliability test conducted in a field setting.

Design

I will now analyze the reliability test according to the two units of analysis that I mentioned at the start of this lecture. As I said, the first unit was the questions in the questionnaire; this is usual in reliability tests. We used multiple regression to explain any variance in the reliability of the questions, by characteristics like the source of information or the area of care that the question addressed. Our second unit of analysis was the institution itself. Here we used multiple regression to explain any variance in reliability among institutions, either by characteristics of the test's administration – for example, the time that elapsed between the two inspections – or by institution characteristics, such as size.

First Unit of Analysis: The Question

We defined four independent variables related to the characteristics of the questions in the questionnaire:

- **The source of information.** This referred to the source from which data were gathered during inspection.

- The **agent of information**. The member of the inspection team who collected the information: either a social worker or a nurse.
- The **area of care** examined. We defined three areas of care: (1) nursing care; (2) social services; and (3) structure.
- The **clarity of the question**. Three experts on questionnaires of this type were asked to indicate whether the question was (1) clear and could be interpreted in only one way, or (2) unclear and could be interpreted in more than one way

RESULTS

The average I_T was .70. The average agreement proportion was .77, and the average kappa was .30. The correction for chance agreement of I_T is more moderate than that of kappa. Moreover, the correlation between I_T and the agreement proportion was .95, while the correlation between kappa and the agreement proportion was .20. This reflects how far kappa is from the intuitive measurement of agreement.

In this Table we can see the results of the multiple regression analysis. The independent variables explain 24 percent of the variance among the questions, as measured by the I_T coefficient. For the variable “source of information”, the categories “general interviews with staff” and “observation” had higher I_T than did the category “interviews with residents”, which was left out of the regression.

For the variable “agent of information”, the social worker had a higher I_T than did the nurse. For the variable “clarity of the question”, “clear” questions had a higher I_T than did “unclear” ones.

Second Unit of Analysis: The Institution

The dependent variable was the institution agreement proportion. It represented the proportion of agreement between the two inspectors regarding **a given institution**. It was expressed as follows:

$$A' = F_0 / \text{TOTAL}$$

where F_0 was the number of items for which there was **agreement**; and TOTAL was the total number items examined. Therefore, the institution agreement proportion ranged from 0 to 1.

There were five independent variables

- **The inspectors who administered the reliability test.**
- **The time that elapsed between inspections.**
- **Quality of care**
- **The size of the institution.**

RESULTS

We analyzed this unit separately for questionnaires completed by social workers and questionnaires completed by nurses.

For all of the institutions, the average institution agreement proportion was .78 for questionnaires completed by social workers, and .68 for questionnaires completed by nurses.

For the social work questionnaires, the independent variables explained 71 percent of the variance in the institution agreement proportion.

For the nursing questionnaires, the independent variables explained 59 percent of the variance in the institution agreement proportion. It appears that in the social worker questionnaires, but not in the nurse questionnaires, part of the variance is a result of different surveyors having completed the questionnaires during the first and second tests. We found that the time that elapsed between the first and second administrations of both types of questionnaire was not significant in explaining the variance in institution agreement proportion.

The size of the institution did not explain the variance in the institution agreement proportion. The quality of care, especially in the nurse questionnaires affected dramatically the variance in institution agreement proportion. In that case, the better the quality of care, the more reliable the questionnaire proved to be.

DISCUSSION

First Unit of Analysis: The Question

On average, the questions reached the level of .70 suggested by Nunnally and others as a benchmark for reliability. Since this study examined an early version of the tool, this average reliability appears to be reasonable. Nevertheless, 39 percent of the questions did not reach this benchmark. Consequently, in later versions of the questionnaires, these questions were dropped or changed. Based on our experience, I will try to show how the results of a reliability test used to improve the tool being evaluated.

The source of information that was found to be least reliable was the interviews with institution residents. The complexity of the interaction with them during an interview affects

responses. For example, a respondent may answer the same question differently at different times, depending on his mood. In addition, different interviewers have different styles.

An interviewer may elicit different answers to the same question, depending on how he asks a question, his tone of voice, or his body language. An interviewer's personal characteristics, such as age and gender may also influence how a question is answered. For example, if the first interviewer was more empathetic than the second, a resident might have answered intimate questions more candidly during the first interview than during the second.

Interviewing elderly people is especially difficult, because they may suffer from impaired hearing, confusion, and fatigue. Moreover, because it can be difficult to communicate with elderly people, an inspector must be flexible in conducting the interview. For example, he may wish to explain the questions rather than reading them as written. When this is the case, however, the measurement tool may no longer be identical every time it is used.

One way of dealing with such problems is to simplify all of the questions, thereby reducing the need for "flexible interviewing". Questions should be short and clear, and use simple terms. However, if an explanation is necessary to exclude multiple interpretations, it should be included in the question.

For example, our tool contained one question for residents with vision impairment: "Have you ever received guidance with orientation to your surroundings?" "Guidance" is an abstract concept. Revising the tool, we change this into several questions as: "Have you ever been taken on a personal tour of the institution? Were obstacles pointed out to you? Have you

received an explanation of how to use the elevator? Were you then taken on a second tour, where you were the leader and the nurse followed you?”

This example illustrates what Lazarsfeld calls “the principle of specification”. In other words, any question that is open to several interpretations should be specified. This will ensure that there is only one legitimate interpretation for each question.

Another principle, which we usually adhered to in these questionnaires, involves using dichotomous answers rather than a multiple choice format, which is complex cognitively, and can tire the elderly respondent. It is also possible to preclude fatigue by dividing the interview into two parts, with a break in between them.

We also wanted to understand why questions administered by nurses proved to be less reliable than questions administered by social workers. One reason may be the intimate nature of some nursing questions, such as questions about urinary incontinence. To help a resident to truthfully answer intimate questions, a nurse must be very empathetic. It might also be wise for her to approach intimate questions gradually.

For example, she might begin by asking “neutral” questions, such as questions about the distribution of medications, and gradually ask increasingly intimate questions, like those about problems with vision, mobility, washing, and, ultimately, incontinence.

Second Unit of Analysis: The Institution

Our analysis of reliability using the institution as the unit of analysis complemented our analysis of reliability. We used the findings to better understand the environment in which the test was carried out.

The use of different inspectors had a significant effect on the variance among institutions. We minimized the variance among inspectors by training them in how to gather information and how to interpret guidelines and regulations.

It appears that the tool is equally reliable in large institutions as in small institutions, even though the surveillance of a large institution places a heavier burden on the inspector.

Both types of questionnaires proved to be less reliable in poor-quality institutions than in better-quality institutions. It is possible that this is because the inspector of a poor-quality institution needs to exercise more judgment to discern whether an item is “deficient” or “not deficient”.

As I have tried to show, the effectiveness of an interrater reliability test is not limited to the single item level only. Rather, multi-variable procedures can enrich our understanding of the tool being tested as well as of the environment in which the reliability test took place.