J Chron Dis 1978, Vol. 31, pp. 573–580.

0021-9681/78/1001-0573$02.00/0

INTER-OBSERVER RELIABILITY OF A FUNCTIONAL STATUS ASSESSMENT INSTRUMENT*

ALAN M. JETTE† and O. LYNN DENISTON

†Department of Health Behavior and Health Education, School of Public Health,
‡Department of Health Planning and Administration,
The University of Michigan, 1420 Washington Heights, Ann Arbor,
MI 48109, U.S.A.

(*Received in revised form* 15 *March* 1978)

Abstract—Using both a concordance and intra-class correlation coefficient approach to assessing inter-observer reliability, the Pilot Geriatric Arthritis Project (PGAP) functional assessment form demonstrates a respectable degree of reliability. The greatest degree of reliability occurs in the dependence dimension, the lowest degree occurs in the difficulty dimension while the degree of reliability for pain ratings falls in between the two.

The PGAP functional assessment instrument has greater reliability when used to rate clients who perform at a relatively high level of functional status as was the case for the general population of PGAP. However, reliability decreases as functional status decreases, at least until total dependence is reached. In other words, the instrument is less reliable when used to score clients with greater degrees of dependence, difficulty, and pain. In the latter case, the lack of concordance may be due to the way in which questions are asked, the way the client responds, the way an interviewer interprets an answer or some combination of these causes. Further work on training and standardizing interviewers is indicated for future studies using this assessment instrument.

INTRODUCTION

The Pilot Geriatric Arthritis Project (PGAP)*, was developed to test the hypothesis that a multidisciplinary health team could improve the 'quality of life' of older adults with arthritis. The project attempted to achieve this aim utilizing current technology of arthritis management through the coordinated delivery of optimum levels of services to clients in the least intensive care setting. Project objectives included prevention of disability, physical restoration, relief of pain, and personal and emotional adjustment. During the PGAP's 3 years of existence, 1089 clients were served. The details of that program have been described elsewhere [1].

Two types of evaluation were planned—measures of 'process' and of 'outcome'.

This paper describes the 'outcome' measures developed for the PGAP; that is, the extent to which project objectives were achieved, and the results of pilot efforts to study their reliability. Subsequent reports will discuss validity of these measures and attempts to assess process.

METHODS

The PGAP staff believed that 'outcome' (i.e. degree of non-disability, physical function, and personal and emotional adjustment) could best be evaluated by assessing the 'functional status' of the project clients. The PGAP staff had been impressed by the work of Katz [2, 3] on developing measures of 'activities of daily living' (ADL) and this guided the search for a measure.

* This project was supported by a contract, administered by the Regional Medical Programs.

In the 'patient classification approach' (PC), an individual's 'functional status' is assessed by measuring the degree of assistance an individual requires to perform 14 different ADL;

Mobility	Stair climbing	Toileting
Transferring	Bathing	*Bowel function
Walking *Wheeling	Dressing	*Bladder function
	Eating/feeding *Behavior pattern	Communication of needs *Orientation as to time, place, person

Degree of dependence in the usual performance of ADL can range from: 0—independent, 1—uses mechanical equipment, 2—uses human assistance, 3—uses both, to, 4—can not perform the activity; as an individual's 'score' increases, his/her degree of dependence also increases. Data can be collected by client self report, professional observation of client performance, or medical chart audits.

The PC contained three limitations that led the PGAP staff to adapt and expand it for use with ambulatory, non-institutionalized adults with arthritis.

Firstly, since the PC was developed for use with institutionalized individuals, five of the 14 categories of ADL were not particularly relevant for the client population of the PGAP: these five categories were eliminated in the PGAP modified version (those items preceded with an asterisk).

Secondly, it was felt that the PC was too narrow in scope and scaled in such a manner as not to account for, or to account inappropriately for, areas of potential functional improvement. For instance, if a client began using a walking aid, a recommendation that was frequently made by PGAP clinical staff, the PC would record that client's status as having 'increased in functional dependence'; it would not account for the potential reduction in pain or difficulty that might result from such a recommendation. To remedy this limitation, the PGAP instrument was designed to include these dimensions in performing an ADL, as well as the 'dependence' dimension in the original PC instrument.

Thirdly, since the PGAP assessment instrument was to be used by the clinical staff as a program planning tool as well as being used as an evaluation tool, the PC was felt to be too imprecise. Since program planning for clients involved a much wider range of ADL, the nine remaining functional categories of the PC were expanded to 44 different functional activities of daily living.

In the PGAP functional assessment instrument, degree of dependence, pain, and difficulty were assessed for each of the specific 44 functional activities. Consequently, the total functional classification could include as many as 132 specific pieces of data (Appendix 1).

In the PGAP instrument, a score for an individual's dependence in an activity was scaled by the same system as the PC. Scores for pain and difficulty were ranked on a four point scale which included 1—no pain/difficulty, 2—mild pain/difficulty, 3—moderate pain/difficulty, and, 4—severe pain/difficulty in performing an activity. For all three dimensions, the data were collected from client self report in a face-to-face interview.

By modifying and expanding the PC in such a manner, it was felt that an *increase* in functional dependency due to health professional intervention would be counteracted by a corresponding *decrease* in functional pain and difficulty.

To facilitate interpreting the data, summary scores were developed by grouping the 44 specific functional items into three general categories labeled 'mobility', 'personal care', and 'work'; these categories contained 12, 17, and 15 items respectively. The items and their assignment to categories are reported in Appendix 1.

One should note that the scales developed in the PGAP instrument are ordinal in nature. However, in collapsing the specific items into categories the data were treated

as interval data and parametric statistics were used; the category score used in the analyses were means of the individual scores for the specific items in each category. The acceptability of this procedure is discussed by Labovitz [4].

Assessment of reliability

A pilot study was initiated by the PGAP staff in the late fall of 1976 to assess the extent of inter-observer reliability in using the PGAP functional assessment instrument as a method of measuring the functional status of adults with arthritis. In this study inter-observer reliability was conceived as the extent to which the same rating or score would be obtained by different interviewers upon assessment of functional status of the same set of clients two or three times on the same day. In other words, provided the functional status of clients does not change, do different observers get the same or similar results?

Nine different interviewers participated in this study; five were members of the PGAP clinical staff, representing the disciplines of nursing, physical therapy, and occupational therapy; one was a trained interviewer from the evaluation staff, and three were trained, non-clinical volunteer interviewers.

Nineteen clients participated in the study; all were over the age of 55, 18 had osteoarthritis, and one had rheumatoid arthritis. In total, 55 independent assessments were completed; 17 clients were assessed by three independent interviewers and two clients were assessed by two independent interviewers.

All 55 assessments were accomplished in face-to-face interviews in the client's home or a community site (e.g. church hall); data were collected by client self report. In order to eliminate variation in functional status due to fluctuations in disease activity, all clients were assessed on the same day. Participants were asked to provide estimates of the average degree of dependency, pain, and difficulty experienced performing each of the 44 functional activities during the previous 2 weeks. Random assignment of clients to interviewers was not possible due to scheduling difficulties, however, we believe no bias was introduced because of this.

RESULTS

Concordance

To be able to compare the extent of inter-observer reliability of the PGAP instrument with the PC form, reliability was first analyzed using a concordance approach. Reliability was determined by calculating an 'agreement ratio' defined as the number of observations per item for which different interviewers agree, divided by the total number of observations made for that item.

It is possible to make several comparisons of the PGAP data, collected by personal interviews with a non-institutional population, with PC data, collected *about* nursing home clients by interviews with staff members [5]. Findings are reported in Table 1 for items with considerable similarity in the two instruments.

Comparison of 'agreement ratios' for comparable dependence ratings in the two studies indicates that, in general, PGAP items which are the more specific, yield higher 'agreement ratios'. Overall concordance for dependence ratings between interviewers using the PGAP instrument is 85% as compared to 79% in the PC study.

'Agreement ratios' between interviewers for degree of difficulty and pain in performing functional activities are reported in Table 2; these ratings are considerably lower than those in the dependence dimension. The highest degree of concordance attained in any of the 'mobility' category items for difficulty was 68%; three of the mobility items demonstrated less than a 50% concordance rate. Concordance for degree of difficulty on 'personal care' items yielded a higher rate of agreement with three items attaining 90% concordance or more.

TABLE 1. AGREEMENT RATIOS IN TEST–RETEST RELIABILITY STUDIES USING THE PC AND PGAP FUNCTIONAL ASSESSMENT INSTRUMENTS

PC Class	PGAP Item	PC Study (*N*-56) Dependence	PGAP Study (*N*-53) Dependence
Walking	Walking inside	0.80	0.92
	Walking outside		0.91
Stair climbing	Stairs inside	0.98	0.91
	Stairs outside		0.87
Transfers	Transfer—bed	0.68	0.96
	chair		0.81
	car		0.75
	toilet		0.77
	bath		0.43
Dressing	Dressing—shoes	0.71	0.83
	Hose/pants		0.92
	Shirt/blouse		1.00
Eating/feeding	Eating/cutting	0.82	0.92
	Drinking		1.00
Bathing	Bathing—all areas	0.73	0.67
	Faucets		0.98
Average		0.79	0.85

Reliability of ratings of pain on function are intermediate between dependence and difficulty. On most items agreement ratios are higher than for difficulty, and again, lower for mobility items than for personal care.

On closer examination, the PGAP concordance data suggest that degree of concordance tends to decrease as client functional status decreases. Table 3 illustrates degree of concordance between pairs of interviewers for all 10 mobility items for the five levels of dependence [i.e. independent, uses equipment, uses human assistance, uses both equipment and human assistance, cannot do either, and the final category, not applicable or 'not ascertainable' (NA)].

The size of the 'agreement ratio' can be seen in the diagonal of the matrix. Although the number of measurements in some categories is small, these data do suggest that as dependence increases the size of the 'agreement ratio' decreases, from a high of

TABLE 2. AGREEMENT RATIOS IN TEST–RETEST RELIABILITY STUDIES USING THE PGAP FUNCTIONAL ASSESSMENT INSTRUMENT—ANALYSIS OF DIFFICULTY AND PAIN RATINGS

PGAP Item	Difficulty	Pain
Mobility		
Walking inside	0.60	0.49
Walking outside	0.42	0.51
Stairs inside	0.43	0.66
Stairs outside	0.42	0.66
Transfer—bed	0.68	0.89
chair	0.53	0.66
car	0.57	0.62
toilet	0.68	0.72
bath	0.55	0.68
Personal		
Dressing—shoes	0.74	0.94
Hose/pants	0.67	0.88
Shirt/blouse	0.87	0.92
Eating/cutting	0.90	0.88
Drinking	0.92	0.96
Bathing—all areas	0.72	0.78
Faucets	0.96	0.97
Average	0.67	0.76

TABLE 3. CONCORDANCE MATRIX FOR OBSERVER PAIRS—ANALYSIS OF DEPENDENCE RATINGS FOR 10 MOBILITY ITEMS

When 1 rater reported:	N (No. measurements)	Per cent of time that other rater reported						
		Ind.	Equip.	Human	Both	Cannot	NA	Total*
Independent	(588)	[0.88]	0.10	0.01	0	0.01	0.02	1.02
Equipment	(390)	0.15	[0.80]	0.01	0.01	0.01	0.02	1.00
Human assist.	(10)	0.30	0.30	[0.40]	0	0	0	1.00
Both equipment and Human assist.	(12)	0	0.33	0	[0.17]	0.42	0.09	1.01
Cannot perform	(32)	0.09	0.09	0	0.16	[0.50]	0.16	1.01
NA	(28)	0.21	0.29	0	0.04	0.18	[0.29]	1.01
Total	(1060)							

*Total deviates from 1.00 due to rounding errors.

88% in the independence category ($N = 588$) to a low of 17% in the uses both equipment and human assistance category ($N = 12$). (However, the increase to 50% concordance for 'cannot perform' suggests a U-shaped curve rather than a linear relationship.) A similar trend is observed if one constructs a similar data matrix for the 'personal care' and 'work' items.

The relationship between the magnitude of the 'agreement ratios' and level of functional status of clients was also investigated for the degree of pain and difficulty. The trend for the degree of difficulty and pain items for the mobility category is similar to that found for the dependence ratings; that is, as degree of functional pain and difficulty increases, the reliability of the ratings decreases.

Inspection of the remaining data set suggests that this trend is similar for 'personal care' and 'work' items as well as for the mobility items.

Might the relationship between extent of reliability and level of functional status be attributed to the design of the study, that is, could the repetition of the same questions on subsequent interviews on the same day lead to a tendency for the respondent to systematically alter his/her response on the subsequent interview?

To investigate this possibility, ratings of dependence, difficulty and pain were analyzed for systematic differences between interviews 1 and 2, 1 and 3, and 2 and 3, using a sign test. In none of the comparisons was there a significant change in either direction; a different rating on a second or third measure was just as apt to be higher or lower.

INTRA-CLASS CORRELATION COEFFICIENT APPROACH (ICC)

The intra-class correlation coefficient is recommended as a useful estimate of reliability for continuous data [6, 7]. This approach is made possible in this study by collapsing the PGAP items into functional categories and creating scores based on the weights reflected in the coding scheme.

Derived from a random effects analysis of variance model, the intra-class coefficient defines reliability as the ratio of the variance of scores between clients to the total variance of scores between and among individuals for each of the nine functional categories:

$$\rho = \sigma_A^2/(\sigma_A^2 + \sigma_e^2).$$

The intra-class correlation coefficients can range from 0 to 1; the higher (or larger) the coefficient the more differences in scores are due to differences *between clients* rather than differences in scores of the same client; thus, the larger the coefficient, the more reliable the instrument.

Findings for the nine basic scores as generated by the PGAP assessment instrument

TABLE 4. INTER-OBSERVER RELIABILITY AS MEASURED BY INTRA-CLASS CORRELATION COEFFICIENT (ICC)

Dimensions	Category			
	Mobility	Personal care	Work	Average
Dependence	0.83	0.84	0.69	0.78
Difficulty	0.69	0.82	0.32	0.61
Pain	0.72	0.81	0.71	0.75

are reported in Table 4. The average degree of reliability across all items is 0.78 for degree of dependence, 0.75 for degree of pain and 0.61 for degree of difficulty.

In addition, the data from Table 4 illustrate that across all three dimensions the lowest coefficients are generated for the category of work items and the highest coefficients are achieved in the personal care items.

Using the ICC approach, the extent of reliability appears respectable for all categories except degree of difficulty for work items, largely 'cooking' and 'housecleaning' tasks.

We recognize the ICC is a *relative* measure of reliability; the coefficient of reliability is high when there is little variation among different measures of the same client in relation to the amount of variation among different clients. Thus a high coefficient may occur even though there is considerable variation among different measures of the same client when there is much more variation among clients. Conversely, little variation within the several measures of a given client may yield a low coefficient when the variability among clients is small.

We obtained an approximation of *absolute* reliability of the PGAP instrument scores from the 'within subjects variation' of the ICC by inspection of the standard deviation. This measure, in original units of the scale, is a measure of random variability and allows calculation of a confidence interval for the variability of a score.

Three of the standard deviations are in the range of one-third of a point on the five point dependence scale and the four point difficulty and pain scales. Five are in the range of one-eighth to one-quarter of a point and 1 is nearly one half point. The 95% confidence interval for a single score ranges from 0.10–0.16 for dependence in personal care to 0.35–0.58 for difficulty at work tasks.

Differences were also analyzed *between* interviewers for the few cases where both a 'professional' interviewer (the one who collected nearly all follow-up data) and the 'volunteer' interviewers (the ones who collected nearly all intake data) interviewed the same clients as part of our reliability study. This comparison was possible across seven clients where it was possible to compare 58 independent ratings; the professional interviewer assigned a higher score on 28 ratings, exactly the same score for 14 ratings and for 16 ratings a less severe score. This set of differences is likely to occur by chance. By contrasting scores on each of our nine basic ratings, it was found that there were no statistically significant differences between the two types of interviewers. To the extent that there is a true difference not detectable in our sample, it is reassuring to note that on the average, the professional interviewers assigned scores of 0.11 lower, or 'worse' than volunteers. Thus any 'before–after' comparison made would tend to under rather than overestimate client improvement.

DISCUSSION

The PGAP functional status assessment instrument demonstrates its greatest reliability in the dimension of degree of dependence in performing basic activities of daily living.

Of the 17 specific functional items from the PGAP instrument seen in Table 1, 'agreement ratios' for dependence ratings are 75% or greater for all but two items, bath transfers and ability to wash all areas of the body when bathing. Most of the observed discordance, 28% for bath transfers and 25% for the bathing item, can be attributed to variability in interviewer interpretation of whether an individual performs an activity

'independently' or 'uses equipment'. In reviewing the completed interview schedules it was clear that most disagreement was due to different opinions or reports about what constitutes 'uses equipment', (e.g. handrails, anti-slip mats). Interviewers were instructed not to consider as equipment any item that would normally be used by people of the client's age but who did not have arthritis. A firmer decision rule may have reduced the disagreement found in these items.

In comparing the reliability of the PC and PGAP instruments, the PC measure yields greater reliability only for stair climbing. This finding is almost certainly due to differences in the populations studied; few of the nursing home residents in the PC reliability study climb stairs while all but one of the PGAP clients did climb stairs in or to their home. Four of the five transfers specified in the PGAP instrument yield higher concordance ratios than the general transfer category of the PC. The low agreement of 43% for bath transfer was almost entirely due to lack of agreement on the definition of equipment. Overall, the more specific tasks in the PGAP instrument yield higher reliability.

In comparison to the extent of reliability of the dependence dimension, agreement ratios for degree of pain and difficulty on performing functional activities achieved lower levels of reliability. The general lack of agreement in rating degree of difficulty and pain is due in part to client reluctance or inability to use the fixed alternative response categories used to scale those items, (i.e. none, mild, moderate, severe). Answers such as 'a good deal', 'right smart', 'a bit' were frequently encountered. In conducting the interviews, if the interviewer was unsuccessful in persuading the client to use the fixed alternative, the interviewer was instructed to interpret the response that was given into the most appropriate fixed alternative. Less interpretation is needed for degree of functional dependence which is a more concrete concept, but two aspects of the measurement system tend to reduce reliability even there, (1) determination of what constitutes special equipment and (2) communication and interpretation of the concept, 'on the average over the last 2 weeks'.

In looking at reliability at differing levels of dependence, pain and difficulty, we note the differences within functional items. For dependence, we see a hint of a *U*-shaped curve; reliability is higher for people independent or completely dependent, lower for intermediate degrees of dependence. This is suggested both by the high concordance for stair climbing in the PC study and the higher reliability at 'cannot do' then 'needs human assistance', with or without equipment in the PGAP study.

In rating pain and difficulty, we note linear rather than *U*-shaped curves; ratios for 10 difficulty and 10 mobility items were 0.71, 0.37, 0.39 and 0.24 when one rater scored none, mild, moderate, or severe respectively. Further work is needed in improving reliability in these mid ranges for dependence and mid and higher ranges for pain and difficulty, especially when the instruments are used for evaluation of intervention programs.

Whereas relative reliability as measured by the intra-class correlation coefficient may be adequate for assessment of programs, it may not be adequately measured for purposes of individual client assessment. The absolute measure, standard deviation of the within client errors, will be more useful here. In PGAP, this was important for care planning where professionals relied upon the intake assessments to determine if their services were needed for each client.

This measure of 'absolute reliability' can thus be useful in individual care assessment and in helping to determine the sample size needed for adequately reliable estimates in population studies.

The degree of reliability obtained in this study does not seem to be an artifact of the procedures used. Second and third assessments of the same client on the same day yielded differences in both directions at equal rates; there was not a systematic increase or decrease in scores on subsequent measures in response to prior measures. Further, there is similar reliability of scores obtained by trained volunteers and professional interviewers and clinical personnel.

Acknowledgements—The authors wish to thank the clients, volunteers, and staff of the Pilot Geriatric Arthritis Program for their cooperation in the collection of the data for this study; the authors also thank R. Landis and H. Ripps for their assistance in the analysis of this investigation.

REFERENCES

1. Final Report: Pilot Geriatric Arthritis Program. Submitted to Wisconsin Association for Regional Medical Programs, Sept. 1977
2. Katz S, Ford AB, Downs TD, Adams M, Rusby DI: Effects of Continued Care: A Study of Chronic Illness in the Home. DHEW Publication No. HSM 73-3010, Dec. 1972
3. Jones E, McNitt B, McKnight E: Patient Classification for Long Term Care: User's Manual. p. 33. DHEW Publication No. HRA 75-3107, Nov. 1974
4. Labovitz S: The assignment of numbers to rank order categories. **Am Soc Rev** 35: 515–525, 1970
5. An approach to the assessment of Long Term Care. Final Report of Research Grant HS-01162, Harvard Center for Community Health and Medical Care, Boston, MA., Dec. 1975
6. Landis JR, Koch GG: A Review of Statistical Methods in the Analysis of Data Arising from Observer Reliability Studies—1. **Statistica Neerlandica** Nov. 1975
7. Fleiss JL, Shrout PE: The Effects of Measurement Errors on Some Multivariate Procedures. **Am J Publ Hlth** 67: 1188–1191, 1977

APPENDIX 1

PGAP FUNCTIONAL STATUS ASSESSMENT INSTRUMENT ITEMS

Mobility	Personal care	Work
Driving—other transportation	Using a phone	Employment/occupation
Shopping	Writing	Using stove/oven/refrigerator
Walking inside	Cutting food	Using sink/faucets
Walking outside	Drinking	Reaching cupboards (high/low)
Stairs in/to home	Ability to wash all areas	Lifting pots/pans
Other stairs	Turning faucets	Peeling/cutting
Curbs	Teethcare	Opening containers
Transferring to/from bed	Shaving	Doing the laundry
Transferring to/from chair	Combing hair	Sweeping/mopping
Transferring to/from car	Washing hair	Bedmaking
Transferring to/from toilet	Setting hair	Washing dishes
Transferring to/from bath	Putting on shoes and tying	Cleaning bathroom
	Putting on hose/pants	Washing windows
	Putting on underclothes	Doing home repairs
	Putting on shirt/blouse	Doing yardwork
	Putting on buttons/zippers	
	Putting on sweater/coat	