Journal of Intellectual Disability Research

VOLUME 42 PART 1 pp 22–28 FEBRUARY 1998

A reliability study of the Spanish version of the Social Behaviour Schedule (SBS) in a population of adults with learning disabilities

L. Salvador-Carulla,[1] M. J. Garcia-Mellado,[1] R. Velazquez,[1] C. Romero[1] & F. Alonso[2]

1 *Grupo de Investigacion en Medicina Psicosocial, Universidad de Cadiz, Cadiz, Spain*
2 *Asociacion PROMI, Cabra, Spain*

Abstract

The reliability of the Spanish version of the Social Behaviour Schedule (SBS) was tested in a vocational setting on a sample of 64 subjects with learning disabilities. Test-retest assessment showed a good percentage of agreement (80%) and adequate kappa values for most SBS items. The overall percentage of agreement of inter-rater reliability was 85% and kappa values were moderate to nearly perfect for 52% of items. Inter-informant analyses produced poorer results, with an average agreement of 43% and inadequate kappa values on 42% of items. The intraclass correlation coefficient (ICC) was 0.64 for test-retest, 0.76 for inter-rater assessment and 0.94 for inter-informant assessment. The Spearman correlation coefficient was adequate on the test-retest and inter-rater analyses, but not on inter-informant analysis. This low inter-informant agreement could be attributed to environmental factors which alter the reliability of reports from different informants in community settings with high levels of normalization. In such environments, an interview with a key informant may not suffice, and both a careful review of the clinical record and a direct interview with subjects may enhance the reliability of the information attained.

Correspondence: Professor L. Salvador-Carulla, Centro de Investigacion de Minusvalias, Universidad de Cadiz PROMI, Urbanizacion El Bosque, Edificio Cycas, 7C, 11405 Jerez de la Frontera, Cádiz, Spain.

© 1998 Blackwell Science Ltd

Keywords Social Behaviour Schedule, reliability, learning disabilities, mental retardation

Introduction

The American Association on Mental Retardation?s classification manual for intellectual disability (AAMR 1992) stresses the need for valid and reliable instruments to evaluate maladaptative behaviour. However, there is a lack of international consensus on the conceptual framework, terminology and standard codification of these conditions. There is also a need for further studies on the psychometric properties of the instruments available in this area, particularly on their transcultural reliability.

The present study aims to evaluate the reliability of the Spanish version of the Social Behaviour Schedule (SBS) (Wykes & Sturt 1986) in an adult population with learning disabilities. The SBS has been used extensively in chronic psychiatric populations, particularly in the assessment of deinstitutionalization and the implementation of community programmes. Some of these studies included individuals with learning disabilities (Hallam *et al.* 1995). A modification of the SBS has been used for assessing deinstitutionalization and community living of people with learning disabilities in England (Cambridge *et al.* 1994). Unfortunately,

there is a lack of information regarding the psychometric properties of SBS in this particular population.

The Social Behaviour Schedule (SBS) is an instrument for assessing behavioural problems based on the work of Lorna Wing with institutionalized psychiatric patients. The inventory has 21 items and the same number of behavioural areas. These areas cover a series of problems which researchers consider to be the most frequent in chronic psychiatric patients, interfering with their daily functioning and their social and family life. Using a five-point Likert scale, these items evaluate the subject's ability to relate to others, the adequacy of her/his social behaviour, and her/his adaptation to environmental requirements. This information refers to the patient's conduct over the past month, as derived from a semi-structured interview with a key informant (i.e. a staff member). In addition to the descriptive profile, two kinds of scores can be obtained from the inventory. The Mild and Severe Behaviour Problems Score (BSM) quantifies the number of behavioural problems which the patient presents with. It is equal to the number of items with a score of two or more. The Severe Behaviour Problems Score (BSS) reflects the number of behaviours which the patient presents that constitute a severe problem. It is equal to the number of items with a score of three or four.

The SBS has shown good inter-rater, test-retest and inter-informant reliability indices (Wykes & Sturt 1986). Studies have been conducted on the reliability of the Spanish version (Vázquez & Jimenez 1989) and its concurrent validity with the Life Skills Profile (Fernandez de Larrinoa *et al.* 1992) in a chronic psychotic population. Unfortunately, these Spanish studies did not include inter-informant assessment in their reliability analyses.

Methods

Procedure

The study was conducted using a working population of subjects with intellectual disability from the Cabra centre of the Asociación para Promoción del Minusválido (Association for Promotion of the Handicapped, known by its Spanish acronym, PROMI). This non-governmental organization has several centres all over Spain. Most of the subjects at this centre come from psychiatric hospitals and institutions. PROMI also provides accommodations, both in residences and the community, for its employees.

A sample of 64 individuals was selected from the 130 subjects with intellectual disability in the Cabra vocational programme at the time of the study. In order to increase clinical variability for the agreement analysis, the study sample included 42 subjects selected randomly from the total population and 15 individuals with behavioural problems placed in a special workshop for people with behavioural problems. All subjects met DSM-III-R criteria for intellectual disability, which was confirmed via clinical and psychometric assessments by one of the evaluators.

Assessment instruments included: the 1990 version of the Assessment and Information Rating Profile (AIRP) (Bouras & Drummond 1990); the Nivel de Satisfaccion y Apoyo (NSA) (Level of Support and Satisfaction) (Salvador-Carulla *et al.* 1996); and global analogue ratings for clinical status and social functioning.

Three types of reliability analysis were carried out in two different phases: (1) inter-rater; (2) test-retest; and (3) inter-informant. The first and the second analyses were performed on 42 individuals (27 selected randomly, plus 15 from the behavioural workshop). The inter-informant analysis was carried out on 37 subjects (22 selected randomly, plus 15 from the behavioural workshop). Twenty-two of the 37 evaluations conducted corresponded to individuals not included in previous evaluations, since it was difficult to find informants with a similar knowledge of the subjects in the study's environment (affirmative industry). During previous explanatory meetings, informants were trained in using the inventory and the psychopathological concepts involved.

Reliability was evaluated according to a standard methodology (Salvador-Carulla 1996). Inter-rater reliability was assessed by two evaluators, who conducted 42 interviews with a reference informant. To avoid observational variation, each evaluator carried out half of the interviews, and these were performed independently by both evaluators.

Sixteen interviews were videotaped. After 30 days, a single rater carried out a second interview, both

with the individual and the same reference informant, to assess test-retest reliability. Inter-informant reliability testing was conducted during a second phase. The rater interviewed two reference informants who knew the subject well during two different sessions. Both interviews were conducted during the same week.

Statistical analysis

Reliability was assessed using the kappa statistical coefficient (Cohen 1960). The test-retest analysis used was that devised by Fleiss (1981) for two scores from a scale with κ categories and n subjects. The intra-class correlation coefficient (ICC) is another reliability coefficient recommended by Bartko & Carpenter (1976), which has advantages over the others in quantitative scales with more than two evaluators, but its use is questioned when dealing with ordinal scales. Not all studies on behavioural assessment include these parameters. The basic statistical proofs recommended for reliability calculation are the kappa coefficient (Bech *et al.* 1993), and in addition, the confidence interval (CI) to specify test-retest agreement.

In the present study, four statistical indices were obtained for each type of reliability to compare total scores: percentage of agreement; weighted kappa (Wj) (Cohen 1968); Spearman correlation coefficient (Rho); and intra-class correlation coefficient (ICC 1,1) (Bartko & Carpenter 1976). In addition, in the test-retest study, a confidence interval of 95% (95% CI) was found.

In keeping with Feinstein (1985), the following correspondence was used to analyse the kappa agreement results: (poor) $\kappa < 0$; (low) $0.0 \leqslant \kappa \leqslant 0.20$); (fair) $0.21 \leqslant \kappa \leqslant 0.40$; (moderate) $0.41 \leqslant \kappa \leqslant 0.60$; (strong) $0.61 \leqslant \kappa \leqslant 0.80$; and (nearly perfect) $0.81 \leqslant \kappa \leqslant 1.0$). Statview and Excel programs for an Apple Macintosh computer were used for these calculations.

Results

The median age of the sample was 35.87±9.58 years (range 20–59 years). The majority of subjects were both single (98%) and male (75%), and came from institutional centres (65.6%). The level of intellectual disability was mild in 37 individuals (57.8%), moderate in 19 (29.6%) and severe in seven (10.9%). Thirty-three per cent of these subjects presented with a DSM-III-R Axis I disorder at the time of the interview: schizophrenia (6); mood disorder (5); schizoaffective disorder or atypical bipolar disorder (3); adjustment disorder (2); anxiety disorder (1); and disorders not otherwise classified (4).

The SBS showed adequate internal consistency (Cronbach's alpha: 0.88). Out of the 21 problem areas assessed by the SBS, the average rate for this sample was 7.8% (Table 1). Self Care (item 17) showed the highest frequency of mild and severe scores (21.8%), followed by Poor Attention Span (item 20), Slowness (item 18), Little Spontaneous Communication (item 1) and Overactivity Restlessness (item 9). The rate of severe behavioural problems (score of 3 or more) was low for this population. The areas with more frequent severe scores were: Overactivity Restlessness, Poor Self Care (item 17), Panic Attacks and Phobias (item 8), and Underactivity (item 19).

Table 1 shows the distribution of behavioural problems in relation to the associated presence of psychiatric disorder according to the DSM-III-R. These two groups (intellectual disability with and without dual diagnosis) presented no significant differences in the overall mild and severe scores (BSM) (Mann-Whitney U-test). However, in the severe scores (BSS), there was a significant difference in favour of subjects with associated psychiatric diagnosis ($z = -2.5$; $P < 0.01$).

Test-retest reliability

Item-by-item reliability and the total percentage of agreement were high for the present study (see Table 2). The average percentage of agreement was 80.1%, with a range of 50–97.6%. The kappa value was poor for item 3 (Inappropriate Conversation) in spite of a good percentage of agreement.

This discrepancy was caused by high variability in the three positive ratings on this item. Kappa values were fair for items 13 (Socially Unacceptable Habits) and 14 (Violence or Threats). Both items showed a high percentage of agreement, and most discrepancies were situated between the absence of the problem and its presence (with a difference

Table 1 Comparison of the frequency and severity of behaviour problems according to the Social Behaviour Schedule in intellectually disabled subjects with and without psychiatric diagnosis

	Intellectual disability (without psychiatric disorder)				Dual diagnosis			
	BSM		BSS		BSM		BSS	
Items	No.	%	No.	%	No.	%	No.	%
(1) Little Spontaneous Communication	6	13.95	0	0.0	3	14.28	0	0.0
(2) Incoherence of Speech	1	2.32	0	0.0	3	14.28	3	14.28
(3) Odd or Inappropriate conversation	0	0.0	0	0.0	1	4.76	1	4.76
(4) Inappropriate Social Mixing	1	2.32	1	2.32	0	0.0	0	0.0
(5) Hostility	3	6.97	0	0.0	1	4.76	1	4.76
(6) Attention-Seeking Behaviour	3	6.97	0	0.0	4	19.14	2	9.52
(7) Suicidal Ideas or Self-harm	0	0.0	0	0.0	0	0.0	0	0.0
(8) Panic Attacks and Phobias	3	6.97	0	0.0	5	23.8	3	14.28
(9) Overactivity and restlessness	5	11.62	2	4.65	4	19.14	3	14.28
(10) Laughing and Talking to Self	4	9.3	1	2.32	2	9.52	1	4.76
(11) Acting Out Bizarre Ideas	0	0.0	0	0.0	1	4.76	0	0.0
(12) Posturing and Mannerisms	1	2.32	1	2.32	1	4.76	1	4.76
(13) Socially Unacceptable Habits or Manners	3	6.97	1	2.32	0	0.0	0	0.0
(14) Destructive Behaviour	1	2.32	1	2.32	1	4.76	1	4.76
(15) Depression	0	0.0	0	0.0	2	9.52	0	0.0
(16) Inappropriate Sexual Behaviour	0	0.0	0	0.0	1	4.76	0	0.0
(17) Personal Appearance and Hygiene	10	23.25	1	2.32	4	19.04	1	4.76
(18) Slowness	6	13.95	0	0.0	4	19.04	0	0.0
(19) Underactivity	5	11.62	2	4.65	1	4.76	1	4.76
(20) Attention Span	6	13.95	1	2.32	6	28.57	0	0.0
(21) Behaviours NOS*	2	4.65	0	0.0	2	9.52	1	4.76

*Not otherwise specified.
BSM: Mild and severe Behaviour Problem Score.
BSS: Severe Behaviour Problem Score.

range of 1). The Wκ coefficient could not be calculated for item 7 (Suicidal Ideas or Behaviour) and 11 (Acting Out Bizarre Ideas) because of the lack of variability in the scores.

Inter-rater reliability

Table 2 shows the total agreement percentage and Wj for inter-rater reliability. The average percentage of agreement was 84.9%, with a range of 57.1–100% for items 17 (Poor Self-Care) and 7 (Suicidal Ideas or Behaviour), respectively. The Wj coefficient could not be calculated in items 7, 11 (Acting Out Bizarre Ideas) and 16 (Inappropriate Sexual Behaviour) because of the lack of variability in the scores obtained.

Inadequate kappa values were found for four items (19%). Agreement was low for item 13 (Socially Unacceptable Habits). Negative scores were obtained on items 2 (Incoherence of Speech), 3 (Odd or Inappropriate Conversation) and 14 (Violence or Threats). As in previous cases, most of the subjects did not score on these items, and when they did, there were discrepancies on the part of their evaluators.

Inter-informant reliability

The average rate of agreement was 43%. The kappa values obtained did not reach a significant level on nine items (42%). Only six of the items showed a moderate kappa. Negative kappa was obtained on items 11 (Acting Out Bizarre Ideas) and 13 (Socially Unacceptable Habits). Item 11 showed a high percentage of agreement (91.8%) since the majority of subjects did not score on this item; however, on the three occasions when it was scored, there were

Table 2 Social Behaviour Schedule reliability study: (PA) percentage agreement; (WPA) weighted percentage agreement; (Wκ) weighted kappa; and (CI) confidence interval

	Inter-rater			Inter-informant			Test-retest			
Item	PA	WPA	WK	PA	WPA	WK	PA	WPA	WK	95%CIWK
1	60	79	0.46	65	88	0.43	64	90	0.5	0.37–0.63
2	93	97	–0.02	76	90	0.19	95	98	0.48	0.35–0.61
3	88	96	–0.04	84	95	0.32	93	97	–0.02	–0.01–0.05
4	74	90	0.37	62	88	0.05	79	93	0.55	0.42–0.68
5	88	97	0.48	86	97	0.53	67	89	0.32	0.19–0.45
6	88	97	0.63	70	91	0.39	74	92	0.49	0.36–0.62
7	100	100	†	97	99	†	95	98	†	—
8	90	96	0.45	51	81	0.01	69	88	0.46	0.33–0.59
9	83	96	0.55	68	88	0.51	79	90	0.58	0.45–0.71
10	83	95	0.69	84	95	0.55	83	90	0.64	0.51–0.77
11	98	99	†	92	97	–0.03	98	99	†	—
12	95	97	0.65	76	89	0.29	88	95	0.5	0.37–0.63
13	83	92	0.17	51	87	–0.08	71	89	0.29	0.16–0.42
14	88	96	–0.04	95	97	0.31	81	92	0.28	0.15–0.41
15	88	95	0.35	65	90	0.05	86	95	0.53	0.46–0.66
16	93	98	†	86	94	0.13	93	98	0.55	0.42–0.68
17	57	84	0.45	62	86	0.48	50	84	0.47	0.34–0.6
18	83	96	0.53	62	86	0.06	76	93	0.6	0.47–0.73
19	79	93	0.25	57	80	0.16	69	89	0.5	0.37–0.63
20	81	94	0.62	59	86	0.3	79	93	0.68	0.55–0.81
21	90	93	0.61	84	89	0.56	95	98	0.94	0.89–0.99

*SBS items described in Table 1.

†Lack of variability in the scores.

discrepancies among the informants. The Wκ coefficient could not be found for item 7 (Suicidal Ideas or Behaviour) because of lack of variability in scores.

Intraclass correlation coefficient (ICC) and Spearman correlation coefficient (Rho)

The ICC was 0.64 for test-retest, 0.76 for inter-rater assessment and 0.94 for inter-informant assessment. The Spearman coefficient was found for each of the items in the three reliability studies. This coefficient was adequate for all of the items in the inter-rater and test-retest reliability studies. However, in the inter-informant reliability study, no significant correlations were found for five items: 4 (Inappropriate Social Mixing), 8 (Panic Attacks and Phobias), 13 (Socially Unacceptable Habits or Manners), 15 (Depression) and 18 (Slowness).

Discussion

The SBS showed adequate test-retest and inter-rater reliability in a vocational setting. These results are comparable to those obtained in Wykes & Sturt's (1986) original study and in the Spanish version assessed in a psychotic population (Vázquez & Jimenez 1989). However, the inter-informant assessment data in the present study are less satisfactory. This could be attributable to various causes. Contrary to what occurs in a closed institution, the subject interacts with different social groups which are not closely related (at work and at home, during leisure and free time) in an integrated environment, and may behave differently with each one. In addition, the number of caretakers and supervisors is notably lower than in closed institutions. Therefore, it is not easy to find several informants who know the subject well in each of

these different environments and domains. Although the PROMI association is responsible for the work centres and integration in the community, so that the team supervises the whole process, it was difficult to find personnel with a similar knowledge of the subjects in the study, and this could be related to the low level of agreement found in inter-informant assessment.

This low inter-informant reliability could also be attributed to staff members' inability to fully understand certain items on the questionnaires.

During the preliminary training period, problems in understanding the psychopathological items included on the SBS became clear. It should be kept in mind that personnel in a vocational environment generally have lower qualifications than those in a clinical environment. This problem is related to the construction of the instrument itself, which has a major psychopathological component. Preliminary training may have lessened this problem, and if this were truly a fundamental factor, the present study's test-retest reliability would have been low as well.

A third reason worth considering is related to the instrument's construction. The possibility of terminological problems has already been noted; this could make comprehension more difficult for less-qualified personnel. In fact, the viability of an instrument aimed at informants with very different technical qualifications (doctors, nurses, teachers, caretakers and relatives) is questionable. This is the case of the SBS and other similar scales, such as the Disability Assessment Schedule (DAS) (Holmes *et al.* 1982), in normalized environments. It is also important to note the discrepancy between ICC and kappa values in this case, which adds further questions to the use of ICC in ordinal measures.

Although the inter-rater and test-retest reliability of the SBS showed good results, the inter-informant reliability was low in a highly qualified vocational centre. This can probably be attributed more to environmental characteristics than to other factors. To confirm this hypothesis, similar studies should be performed on larger samples, using MR subjects in clinical settings, as well as on severely ill psychiatric patients in vocational environments. However, it is important to be aware of the fact that a single key informant may provide unreliable information on the behavioural problems shown by individuals with learning disabilities living in the community. In such circumstances, other information sources, such as clinical records and direct patient interviews, should be considered when completing this rating scale.

Acknowledgement

This study was supported by Grant CICYT SEC 92–0862 from the Comisión Interministerial de Ciencia y Tecnología, Spain.

References

American Association on Mental Retardation (1992) *Mental Retardation: Definitions, Classifications and Systems of Support*, 9th edn. American Association of Mental Retardation, Washington, DC.

Bartko J. & Carpenter J. R. (1976) On the theory of reliability. *The Journal of Nervous and Mental Disease* **163,** 307–17.

Bech P., Malt U. F., Dencker S. J., Ahlfors U. G., Elgen K., Lewander T., Lundell A., Simpson G. M. & Lingjaerde O. (1993) Scales for assessment of diagnosis and severity of mental disorders. *Acta Psychiatrica Scandinavica* **87 (Suppl. 372),** 1–87.

Bouras N. & Drummond C. (1990) *Assessment and Information Rating Profile (AIRP)*. Research, Evaluation and Development Unit, Section of Mental Handicap, Division of Psychiatry, United Medical and Dental Schools, Guy's Hospital, London.

Cambridge P., Hayes L., Knapp M., Gould E. & Fenyo A. (1994) *Care in the Community: Five Years On. Life in the Community for People with Learning Disabilities.* Arena Ashgate Publishing Ltd, Aldershot.

Cohen J. (1960) A coefficient agreement for nominal scales. *Educational and Psychological Measurement* **20,** 37–46.

Cohen J. (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* **70,** 213–20.

Feinstein A. R. (1985) *Clinical Epidemiology.* W. B. Saunders, Philadelphia, PA.

Fernández de Larrinoa P., Bulbena A. & Dominguez A. I. (1992) Estudio de fiabilidad, validez y consistencia interna de la escala LSP (Life Skills Profile). Perfil de habilidades de la vida cotidiana. *Actas Luso Españolas de Neurología y Psiquiatría* **20,** 71–5.

Fleiss J. L. (1981) *Statistical Methods for Rates and Proportions.* John Wiley & Sons, Chichester.

Hallam A., Knapp M., Beecham J. & Fenyo A. (1995) Eight years of psychiatric reprovision: an economic evaluation. In: *The Economic Evaluation of Mental Health Care* (ed. M. Knapp), pp. 103–24. Arena, Ashgate Publishing Ltd, Aldershot.

Holmes N., Shah A. & Wing L. (1982) The Disability Assessment Schedule: a brief screening device for use with the mentally retarded. *Psychological Medicine* **12,** 879–90.

Salvador-Carulla L., Garcia-Mellado M. J., Romero C. & Montero S. (1996) *Evaluación Psicosocial del Retraso Mental.* Ministerio de Asuntos Sociales, Instituto Nacional de Servicios Sociales, Madrid.

Salvador-Carulla L. (1996) Assessment instruments in psychiatry: description and psychometric properties. In: *Mental Health Outcome Measures* (eds G. Thornicroft & M. Tansella), pp. 189–206. Springer-Verlag, Berlin.

Vázquez A. J. & Jimenez R. (1989) Problemas de conducta y esquizofrenia: estudio mediante el SBS (Social Behaviour Schedule) en el distrito de Ecija. *Actas Luso Españolas de Neurología Psiquiatrica* **17,** 281–93.

Wykes T. & Sturt E. (1986) The measurement of social behaviour in psychiatric patients: an assessment of the reliability and validity of the SBS. *British Journal of Psychiatry* **148,** 1–11.

Received 17 April 1996; revised 4 June 1996