SOCIAL BEHAVIOR AND PERSONALITY, 2004, 32(6), 535-550

STABILITY OF THREE METHODS FOR TWO-DIMENSIONAL SOCIOMETRIC STATUS DETERMINATION BASED ON THE PROCEDURE OF ASHER, SINGLETON, TINSLEY AND HYMEL

GERARD H. MAASSEN
Utrecht University, The Netherlands
HENDERIEN STEENBEEK AND PAUL VAN GEERT
University of Groningen, The Netherlands

This study aimed at comparing the stability of three methods for two-dimensional sociometric status determination, including (1) the recently developed SSrat technique (Maassen, Akkermans, & Van der Linden, 1996), as well as (2) the procedure of Howes (1988), which is based on the algorithm and classification criteria of Coie and Dodge (1983), and (3) a procedure applied by Cillessen, van IJzendoorn, Van Lieshout, and Hartup (1992) based on the algorithm and criteria of Newcomb and Bukowski (1983). The latter two techniques can be regarded as nomination methods eliciting unlimited numbers of nominations from the participants. In the present study, 83 pupils (ages ranging from 5.8 to 8.8 years) participated in the 3-point rating procedure of Asher, Singleton, Tinsley, and Hymel (1979). They were asked to rate how much they liked to play with each of their classmates. The ratings data were processed in three ways according to the methods mentioned above. The techniques of Howes and of Cillessen et al. showed higher stability than previously ever reported using the nomination procedures of Coie and Dodge or Newcomb and Bukowski. Nevertheless, the rating scale procedure SSrat proved to have even higher stability.

Gerard H. Maassen, Department of Methodology and Statistics, Faculty of Social Sciences, Utrecht University, The Netherlands; Henderien Steenbeek and Paul van Geert, Department of Psychology, University of Groningen, The Netherlands.
Appreciation is due to reviewers including: Philip Erwin, PhD, The Psychology Unit, Edge Hills College, St Helens Road, Ormskirk, Lancashire I39 4QP, England, Email: <erwinp@edgehill.ac.uk>; David Fitzgerald, PhD, Duke University Medical Centre, Duke Child and Family Study Centre, 718 Rutherford Street, Durham NC, US 27705, Email: <fitzg010@mc.duke.edu>; Frits Goossens PhD, Faculty of Psychology and Pedagogy, Van der Boechorststraat 1, 1081 BT Amsterdam, The Netherlands, Email: <fa.goossens@psy.vu.nl>
Keywords: sociometric status determination, nomination, ratings, test-retest reliability, stability, preschool children.
Please address correspondence and reprint requests to: Gerard H. Maassen, Department of Methodology and Statistics, Faculty of Social Sciences, Utrecht University, P.O. Box 80140, 3508 TC Utrecht, The Netherlands. Phone: +31 30 253 4765; Fax: +31 30 253 5797; Email: <g.maassen@fss.uu.nl>

The application of sociometry has expanded enormously during the past decades. At present, researchers prefer to use the nomination methods proposed in the early eighties, particularly the one by Coie, Dodge and Coppotelli (1982), hereafter referred to as CDnom, and to a lesser extent the one by Newcomb and Bukowski (1983), here referred to as NBnom. These methods owed their popularity in part to the fact that they met a need expressed by earlier authors (Dunnington, 1957; Gronlund & Anderson, 1957), namely, the opportunity to classify children who receive few positive nominations into two groups: (1) those who also receive few negative nominations (the neglected category), and (2) those who also receive many negative nominations (the rejected category). These methods can be regarded as two-dimensional since they classify the children along two dimensions: the social preference dimension (the popular, the average and the rejected category) and a dimension labeled *social impact* or *visibility* (the neglected, the average and the controversial category). Since the inception of these methods, numerous studies have looked at the relationships between sociometric status and various correlates, including academic performance, social competence and aggressive behavior. (For a review of the first decade, see Newcomb, Bukowski, & Pattee, 1993).

Since the nomination methods give the opportunity to classify the participants into the five sociometric status groups mentioned earlier, they overshadowed the rating scale method, despite the fact that various publications (Asher & Hymel, 1981; Oden & Asher, 1977; Thompson & Powell, 1951) showed a greater stability for the rating scores than for nomination scores (such as the totals of the positive nominations received or the negative nominations received). Traditionally, with the ratings methods, only sociometric status could be classified along the popularity dimension. In the case of younger children, the rating scale method continued to be used because nominations given by preschool children have been shown to be unreliable (Bronfenbrenner, 1944). Even with the modification proposed by McCandless and Marshall (1957), where children are asked to point to photographs of their classmates rather than to name their best friends, only moderate average test-retest correlations of nomination scores have been reported (e.g., Hartup, Glazer, & Charlesworth, 1967; Moore & Updegraff, 1964). Asher, Singleton, Tinsley, and Hymel (1979) described a method which is basically to present preschool children with a 3-point rating scale in an adapted format. They asked children to assign pictures of each of their classmates to one of three faces according to how much they liked to play with that person: a happy face, a neutral face, and a sad face. The authors reported that the test-retest reliability of the rating scale measure was superior to that of the nomination measures, a finding confirmed by later authors (Olson & Lifgren, 1988). It should be noted that these comparisons were restricted to one-dimensional rating scale measures of social preference, because, at the time, a

two-dimensional classification was not possible with rating scales. Since its introduction, the method of Asher et al. has often been applied.

Maassen, Akkermans and Van der Linden (1996) developed a rating scale method (SSrat)[1] which did enable the researcher to classify individuals along the two dimensions of sociometric status and to realize a categorization into the same five status groups as do the current nomination methods. In this new method, the information needed for a distinction between the neglected and rejected individuals is derived from data collected by means of a bipolar rating scale with a neutral scale midpoint. This is a promising method (Hymel, Vaillancourt, McDougall, & Renshaw, 2002) since, as will be explained later, it processes more refined information than do the nomination methods; certainly in the way the latter methods are traditionally applied, namely, by asking for a limited number (usually three) of positive and negative nominations.

The present study involved younger children who were asked to rate their classmates on 3-point rating scales. A comparison of the results of SSrat and those of the current nomination methods – CDnom or NBnom – is less meaningful when young children are concerned, since these methods are considered less appropriate in this case. However, Howes (1988) and Cillessen, Van IJzendoorn, Van Lieshout, and Hartup (1992) conceived and applied variants which are highly relevant in the present context. Using the method of Asher et al. (1979), Howes collected 3-point ratings among preschool children and applied the algorithm of Coie and Dodge (1983) to the total number of positive ratings received (the acceptation score) and the total number of negative ratings received (the rejection score). This procedure has already been described for an arbitrary number of scale points by Maassen et al. (1996) and was then referred to as *CDCrat*, which name will be used hereafter in the present article. On the other hand, Cillessen et al. (1992) applied an adaptation of the Newcomb and Bukowski procedure to the acceptation and the rejection scores. These two variants are interesting because they can be regarded as respectively a quasi-CDnom or a quasi-NBnom procedure where there is no constraint on the numbers of positive or negative nominations requested. Terry (2000) has shown that abolishing this constraint improves the performance of the CDnom procedure, and it is expected that researchers in the field will increasingly follow this advice to use unlimited nominations in the future. Thus, the quasi-nomination procedures turn out to be of topical interest, and particularly relevant to a comparison with the SSrat "pure" rating scale method. In the present study, attention was directed particularly to the stability of the three procedures.

[1] The computer program and manual (Maassen & Landsheer, 1996) can be requested from the first author.

Stability and Reliability of Sociometric Methods

The *stability* of a given attribute is usually defined as *the correlation between the measurements of the same sample on two occasions*. Where the stability of sociometric classification is concerned, the agreement between two classifications must be expressed by a measure of association for nominal variables, for which Cohen's κ is traditionally used. Cillessen, Bukowski, and Haselager (2000) reviewed the dozen studies that reported the stability of sociometric classifications. The values of Cohen's κ varied strongly, from .01 to .44 (the latter was found by Newcomb and Bukowski, 1983, applying CDnom to assessments with a 1-month interval). This is not surprising, since the stability strongly depends on the social context of the assessments. The conditions under which the data were collected could easily have changed, even within a short period. The composition of subgroups of the total research sample could have been rearranged (for example, class groups within a school). But social preferences may also easily change within fixed subgroups, particularly when young children are involved (Wu, Hart, Draper, & Olsen, 2001).

A different factor affecting the stability is of a technical nature, namely the reliability of the procedure in question. Although the concept of reliability in the context of sociometry will not seem strange to researchers, calibration of reliability with regard to sociometric status determination is problematic. Reliability is a concept originating from psychometrics and applies to interval level measurement instruments of relatively stable traits. The *reliability coefficient* is defined as *the correlation between two measurements within the same sample under similar conditions*. Regarding the two measurements it is assumed that the measurement errors are nonsystematic and not correlated with the "true scores". Furthermore it is assumed that the variances of the observed variables are invariant across assessments; the same applies to the error components of the observables. Under these assumptions the correlation (and thus the reliability coefficient) equals the percentage true variance as a component of the observed variance. The reliability strongly depends on the characteristics of the research sample in question and its value can be generalized only if the sample represents a larger population. This conceptualization cannot be easily transposed into the sociometric context. Sociometric classification is not measured at the interval level. Agreement between two classifications can be assessed, but the concepts of error variance and percentage true variance are meaningless within the sociometric context. Moreover, few researchers will view peer-assessed sociometric status as a stable trait and stability coefficients observed within a study as easily generalizable to larger populations. Nevertheless, Wu et al. (2001) rightly note that to date, in the sociometric literature, little attention has been paid to the difference between measurement reliability and stability. Their complaint applies especially to studies that examine

the quality of different procedures (e.g., Frederickson & Furnham, 1998; Terry & Coie, 1991), and that typically compare κ-values suggesting that the highest value corresponds to a superior performance.

In their study of children of preschool age, Wu et al. (2001) did distinguish between stability and reliability. However, they confined themselves to establishing the reliability of variables indicating the concept *popularity*, but did not report correlations between two consecutive sociometric classifications. In the present study we did both: concordance between classifications across time was reported regarding the procedures involved, but we also attempted to collect indications that the assessments were conducted under similar conditions (regarding the composition of the peer groups, or the social preferences of the group members). In our view, this strengthens our conclusions regarding the stability and provides deeper insight into the reliability of the procedures examined. Only the cross-time stability coefficient can be established in practice and the observed value is always underestimating the reliability coefficent. Differences between the highest stability coefficients found regarding the various procedures are the most trustworthy indications for differences between the reliability coefficients of these procedures. However, as the reader will understand from our discussion, this is not to suggest that, in this way, the reliability of the various procedures is conclusively established.

We note that this study and previous related studies are concerned with global measures of stability. They are based on judgments of a considerable number of individual children who can be considered as conditionally independent of one another. Under this general condition, the distinction – based on psychometrics – between reliability and stability and the distinction between similar and dissimilar measurement circumstances provide a valid framework for research and statistical calculations. However, if attention is directed towards the individual rater, more specifically the stability of individual ratings, a different theoretical framework becomes relevant, namely, that of dynamic systems and contextual processes. For an individual rater, a repeated rating is then considered in the context of the preceding one, and thus, the notion of similar circumstances becomes less relevant. The issue of stability at the individual level is the topic of a study carried out by two of the present authors (Steenbeek & Van Geert, in preparation; Van Geert & Steenbeek, in preparation).

METHOD

Participants

Four class groups originating from three different ordinary elementary schools in the Netherlands participated in the study. One of the schools was characterized by a population consisting of a relatively high number of children from a lower socioeconomic background. Parental consent for participation was obtained for

83 (83.6%) of the total number of 99 pupils, 47 boys and 36 girls. Their average age was 6.5 years, with ages ranging from 5.8 to 8.8 years. The class groups comprised 23, 20, 24, and 16 children, respectively.

Procedure

Pictures were made of all the participating children. One by one the children were taken from their class into a separate room, where, with the help of a laptop computer, the pictures were presented in random order, and the children were asked to rate how often they played with each depicted classmate. The children could choose one of the following anchors: *never, sometimes, often*. Subsequently, the children were asked to rate how much they liked to play with a depicted child, the alternatives being *not liked, ordinary, liked*. Before the actual assessment took place, the child could practise answering the questions until he or she understood the procedure. The procedure was carried out three times. The first assessment was conducted in February/ March 2000, the second in April 2000, and the third in May/ June 2000. The interval between two assessments was minimally 35 and maximally 60 days.

Sociometric Measures

The negative frequency-of-playing rating and the negative like-to-play rating were scored 1, the midpoint ratings were scored 2, and the positive ratings were scored 3. From these scores the average ratings received were calculated, hereafter referred to as *ARRF* for the frequency-of-playing data and *ARR* for the liked-to-play data. The raw rating scores were used as input for the sociometric classification methods involved. The like-to-play ratings constituted the principal variables in the present study. The frequency-of-playing ratings were employed merely for validatory purposes.

The Howes Variant of the Standard-Score Method of Coie et al. (CDCrat). Let L be the number of positive ratings received by a person and D the number of negative ratings received. L and D are standardized within the group. Next, these standard scores are transformed into two new variables: *SP* (*Social Preference*) and *SI* (*Social Impact*) calculated as their difference and sum respectively, hence $SP = z_L - z_D$ and $SI = z_L + z_D$. The resulting scores are standardized as well. Finally, the subjects are attributed to five sociometric status categories as follows:
(a) popular, persons with standardized $SP > 1$, $z_L > 0$ and $z_D < 0$; (b) rejected, persons with standardized $SP < -1$, $z_L < 0$ and $z_D > 0$; (c) neglected, persons with standardized $SI < -1$, $z_L < 0$ and $z_D < 0$; (d) controversial, persons with standardized $SI > 1$, $z_L > 0$ and $z_D > 0$; (e) average, all remaining group members.
The Probability Method of Newcomb and Bukowski as applied by Cillessen et al. (1992) (NBrat). Unlike the Coie and Dodge procedure, nonstandardized

numbers of "liked" nominations *L* and "disliked" nominations *D* are processed. In addition to the scores *L* and *D* only the impact score $SI = L + D$ is used. The original procedure of Newcomb and Bukowski prescribes instructing the respondents to nominate a fixed number *a* of "liked" and "disliked" individuals. Then, under the null hypothesis that nominations are randomly assigned, the probability that a group member receives a "liked" nomination from a fellow group mate is fixed: $a/(N - 1)$, *N* being the group size. The same holds for the probability of receiving a "disliked" nomination. Thus, under the null hypothesis, for each group member *L* and *D* are the sums of *N*-1 Bernoulli variables. If one assumes that a respondent will not give a fellow member both a "liked" nomination and a "disliked" nomination, the impact score *SI* can be seen as the sum of *N*-1 Bernoulli variables too. The probabilities of observed *L*, *D* and *SI* scores are then compared with a chosen significance level. Since, in the present investigation, unlimited nominations were used, these probabilities were computed using the generalized binomial distribution (Ten Brink, 1985). Sociometric status is then attributed as follows: (a) popular, persons with significantly high *L* and *D* less than its expected value; (b) rejected, persons with significantly high *D* and *L* less than its expected value; (c) neglected, persons with significantly low *SI*; (d) controversial, persons with significantly high *L* and *D* greater than its expected value, or significantly high *D* and *L* greater than its expected value; (e) average, all remaining group members.

The Two-Dimensional Ratings System SSrat For an accurate mathematical description of the method applied to rating scales with an arbitrary number of scale points, see Maassen et al. (1996) or Maassen and Landsheer (1998). (A text on an introductory level is provided by Maassen, Van der Linden, Goossens, & Bokhorst, 2000). For 3-point scales the procedure is tantamount to the following. The numbers of the positive and negative ratings received are calculated (*L* and *D*), but not standardized. *SI* (*Social Impact*) is taken as the sum of these two totals, and *SP* (*Social Preference*) as the difference (which is equal to the sum of the ratings received minus a constant). Then, the individual multinomial distribution of the ratings attribution of each assessor is estimated from his or her observed rating pattern. Based on the multinomial distributions, the probabilities of the observed *SP, SI, L* and *D* are calculated. The classification into the five common sociometric status groups is guided by the test of whether these probabilities turn out to be higher or lower than expected on the basis of chance.

RESULTS

Preliminary Analyses

First, we looked at the extent to which the social preferences within the sample had changed during the research by investigating the stability of the assessments

along the social preference dimension. Table 1 already shows very high observed intercorrelations of the average rating received on the three occasions, but the stability should preferably be established corrected for attenuation resulting from measurement error. To this end, the social preference score of the participants was modeled as a latent construct, on each occasion measured by the observed variable *ARR*. The latent and observed variables were cast into a simple structural equations model as depicted in Figure 1. With the plausible assumption that the error variances of the three observables are equal, parameters of the model can be estimated (Wiley & Wiley, 1970), for which we used the LISREL 8 program.

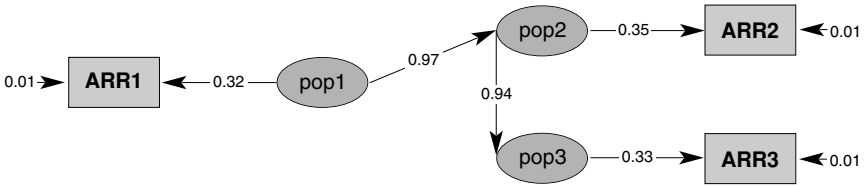

Figure 1: Path model for the stability of peer-perceived social preference measured on three occasions using the Average Like-to-Play Rating (*ARR*) as indicator

The model is saturated, implying that only parameter estimation is possible, whereas model testing is not. The central outcomes are the stability coefficients: 0.97 between the first two assessments, and 0.94 between the second and third assessment.

TABLE 1
AVERAGE LIKED-TO-PLAY RATING RECEIVED AT CONSECUTIVE ASSESSMENTS

time	correlations			*N*	*M*	*SD*	%<midp*	%>midp*
	1	2	3					
1	1.000			84	2.04	0.33	39.3	52.4
2	.870	1.000		85	2.04	0.36	38.8	51.8
3	.815	.845	1.000	82	2.02	0.34	39.0	47.6

Note: * Scale midpoint equals 2.0. The percentages in these two columns do not horizontally add up to 100 when an individual's Average Rating Received exactly equals the scale midpoint.

In addition, we looked at the development of the means of the observed variable *ARR* across the three assessments. Table 1 shows that the mean *ARR* remained practically unchanged and equal to the midpoint of the rating scale (i.e. 2), implying that on average the participants consistently evaluated the popularity of their classmates neutrally.

From the almost perfect stabilities of the social preference concept and from the fact that the mean of *ARR* hardly fluctuated, we inferred that the social

preferences within the class groups remained practically unchanged during the short intervals, and that the stability of the eventual classifications realized in the present study very well reflect the test-retest reliability of the procedures involved.

There is no "gold standard" for the size that popular or rejected categories should have, but most researchers feel comfortable with the outcomes of the standard-score CDnom method, which usually yields relative sizes of 15-20% for these two categories. NBnom and SSrat are probability methods, implying that a value for the significance level α must be chosen. Newcomb and Bukowski (1983) originally proposed to choose $\alpha = .05$, but in practice this generally proves to yield appreciably smaller popular and rejected groups than does CDnom. To achieve groups of comparable size the α value is often increased to .10 (e.g., Newcomb & Bukowski, 1984; Terry & Coie, 1991). In the present study, too, the Newcomb and Bukowski variant using $\alpha = .05$ yielded smaller categories than did CDCrat. Therefore, the Newcomb and Bukowski method was also conducted using $\alpha = .10$. SSrat was applied to the liked-to-play ratings using differing α values, and on all three occasions $\alpha = .05$ proved to yield similar sized popular and rejected categories as did the standard variant of CDCrat.

TABLE 2
SOCIOMETRIC STATUS DISTRIBUTIONS (IN PERCENTAGES) AT CONSECUTIVE ASSESSMENTS ACCORDING TO THREE CLASSIFICATION PROCEDURES. (TIME 1: *N* = 84, TIME 2: *N* = 85, TIME 3: *N* = 82)

time	Method	α	pop	rej	neg	con	ave
1	SSrat	.050	17.9	20.2	1.2	1.2	59.5
	CDCrat		19.0	21.4	11.9	6.0	41.7
	NBrat	.050	11.9	15.5	1.2	0.0	71.4
		.100	16.7	19.0	6.0	0.0	58.3
2	SSrat	.050	23.5	20.0	4.7	0.0	51.8
	CDCrat		21.2	17.6	9.4	3.5	48.2
	NBrat	.050	17.6	15.3	5.9	0.0	61.2
		.100	23.5	20.0	8.2	0.0	48.2
3	SSrat	.050	18.3	19.5	2.4	0.0	59.8
	CDCrat		17.1	19.5	9.8	1.2	52.4
	NBrat	.050	19.5	14.6	2.4	0.0	63.4
		.100	23.2	18.3	11.0	0.0	47.6

CLASSIFICATORY RESULTS

Table 2 presents the outcomes of conducting CDCrat, SSrat with $\alpha = .05$, and the Newcomb and Bukowski variants with $\alpha = .05$ and .10. The Table shows that

SSrat assigned considerably fewer children to the neglected or controversial categories than did CDCrat. (According to SSrat, on the last occasion, nobody at all was designated controversial.) Further analyses of the present data proved that the children designated neglected by SSrat were a stringent selection from the neglected category resulting from CDCrat, being roughly those children with the highest numbers of neutral scores received combined with both a relatively low number of positive and a relatively low number of negative ratings received . Allied to the low numbers of neglected and controversial children, SSrat yielded larger average groups than did CDCrat. The numbers of neglected and controversial children assigned by the Newcomb and Bukowski method using α = .05 resembled those assigned by SSrat using α = .05. The Newcomb and Bukowski method using α = .10 assigned more children to the neglected category than when using α = .05, but this did not hold for the controversial category.

TABLE 3

COMPARISONS BETWEEN THE POPULAR (P) VERSUS THE AVERAGE (A), AND THE REJECTED (R) VERSUS THE AVERAGE GROUPS REGARDING THE AVERAGE RECEIVED FREQUENCY-OF-PLAYING RATING

t	method	α	*N*			average frequency-of-playing rating				
			P	A	R	Mean			ES	
						P	A	R	P v A	R v A
1	SSrat	.050	15	50	17	2.08	1.84	1.46	1.25	-1.84
	CDCrat		16	35	18	2.08	1.83	1.49	1.25	-1.49
	NBrat	.050	10	60	13	2.13	1.84	1.42	1.45	-2.03
		.100	14	49	16	2.11	1.83	1.47	1.49	-1.74
2	SSrat	.050	20	44	17	2.14	1.87	1.51	1.41	-1.71
	CDCrat		18	41	15	2.17	1.85	1.51	1.62	-1.53
	NBrat	.050	15	52	13	2.14	1.88	1.47	1.23	-1.85
		.100	20	41	17	2.14	1.86	1.51	1.48	-1.67
3	SSrat	.050	15	49	16	2.22	1.94	1.52	1.35	-2.07
	CDCrat		14	43	16	2.23	1.95	1.52	1.30	-2.07
	NBrat	.050	16	52	12	2.19	1.92	1.49	1.23	-1.94
		.100	19	39	15	2.19	1.93	1.50	1.29	-2.06

Notes: 1. all *t*-values corresponding with ESs are significant (one-sided $p < .01$).
2. Effect sizes at consecutive assessments, according to three classification procedures.

We carried out a validation analysis with variable *ARRF* as a criterion. Since the distinctive characteristics of the neglected categories are a matter of doubt (Newcomb & Bukowski, 1983; Newcomb et al., 1993), and since the numbers of (particularly) the controversial groups were very small, we confined ourselves to

a comparison of the popular and rejected categories on the one hand and the average groups on the other hand. In Table 3, these contrasts are expressed in terms of effect sizes for the three classification methods on the three measurement occasions. The absolute value of all the ESs exceeded 1, indicating very distinct differences. We note that ESs exceeding .8 are usually regarded as large (Cohen, 1992). The distinction between popular and rejected on the one hand versus average on the other hand across the three measurements constitutes six cases. In three of the cases the ESs were not substantially different (difference between the lowest and highest value .20 or less) while the remaining three cases did not show outcomes in favor of any particular procedure. We conclude that these validation results did not point at one method or variant as being preferable.

Stabilities Compared

Table 4 shows the stability of the two-dimensional classifications resulting from applications of the three procedures in question. The stabilities are expressed in terms of Cohen's κ and calculated for the two consecutive transitions involved (i.e., between assessments 1 and 2 and between assessments 2 and 3). It may be helpful to the reader to cite Landis and Koch's (1977) recommendations for the interpretation of κ values: values less than .30 constitute slight to fair agreement; values between .30 and .60 indicate moderate agreement, whereas values greater than .60 indicate excellent agreement.

TABLE 4
Cohen's Kappa in the Total Sample and in the Separate Status Groups for Two Transitions According to Three Classification Procedures

method	α	pop	rej	neg	cont	ave	total
transition 1-2							
SSrat	.050	.82	.63	-.03	-.01	.54	.61
CDCrat		.71	.74	.25	-.05	.35	.48
NBrat	.050	.48	.64	-.04	--	.34	.43
	.100	.56	.51	.10	--	.21	.37
transition 2-3							
SSrat	.050	.78	.66	.65	--	.58	.66
CDCrat		.64	.80	.31	-.03	.37	.51
NBrat	.050	.59	.58	.55	--	.43	.52
	.100	.59	.61	.31	--	.32	.46

The overall κ values yielded by CDCrat (around .50, see the last column of Table 4) are higher than ever reported for CDnom classifications. The κ values of the Newcomb and Bukowski variant were roughly the same or lower. For both transitions the overall stability of SSrat was even higher than those yielded by

CDCrat: over .60, which is regarded as excellent concordance. The Table also provides information on the stability of the separate status groups. We see that the κ values for the popular and rejected categories resulting from CDCrat and SSrat fluctuated around the same high level. The results of SSrat showed a slightly higher stability of the popular groups than did the results of CDCrat; contrary outcomes were found for the rejected groups. The two Newcomb and Bukowski variants showed lower stabilities for the popular and rejected categories.

The higher overall stability resulting from SSrat is related to differences in the stabilities of the other status categories. SSrat and the Newcomb and Bukowski variant using $\alpha = .05$ designated considerably fewer pupils to the neglected and controversial categories than did CDCrat or the Newcomb and Bukowski variant using $\alpha = .10$. The high instability of the categories is consistently reported in the literature (e.g., see Rubin, Bukowski, & Parker, 1998). In the present study, too, the κ values concerning these two categories proved to be highly variable and were minimal in most cases. (In one case, however, viz. the neglected category produced by SSrat for the second transition the stability was notably high). SSrat consistently assigned more participants to the average category and showed higher stability of the resulting average categories than did CDCrat or the Newcomb and Bukowski variants.

DISCUSSION

The present study involved three assessments within a period of four months. During the short intervals between the assessments the class groups remained practically unchanged, and it could be roughly ascertained that the social preferences had not changed during the course of the research, which leads us to suppose that the differential stabilities of the methods involved fairly reflect the differential test-retest reliabilities. Then, the classificatory results of SSrat were compared with those of the Cillessen et al. (1992) variant of the Newcomb and Bukowski (1983) procedure using $\alpha = .05$ and .10 (here referred to as NBrat), and with the Howes (1988) variant of Coie and Dodge's (1983) procedure (here referred to as CDCrat). Several important differences can be noted between the present research and previous comparative studies: (a) since in the present study nominations and ratings coincide, the procedures involved have been applied to the same data; (b) as far as CDCrat and the Newcomb and Bukowski procedure applied to 3-point ratings can be interpreted as nomination procedures, unlimited numbers of nominations have been elicited from the participants, which is highly relevant to recent insights.

A validation analysis showed that CDCrat, a comparable variant of SSrat (i.e., with $\alpha = .05$), and two variants of the Newcomb and Bukowski method produced popular and rejected categories of practically the same quality (in terms of effect

sizes of the contrasts between the popular and rejected categories on the one hand and the average categories on the other hand). The stability of CDCrat in terms of Cohen's κ was around .50, higher than ever reported for a nomination procedure. At most, the stabilities of the Newcomb and Bukowski method were roughly the same. Nevertheless, the stability of SSrat ($\alpha = .05$) proved to be considerably higher: for SSrat on average (calculated across two transitions) κ was equal to .63, which is regarded as excellent. The difference between the SSrat and CDCrat outcomes can be explained by the following observations: (a) SSrat takes the individual score patterns of the assessors into account, which means the processing of more refined information; (b) SSrat is much more selective in assigning individuals to the neglected or controversial categories. The latter phenomenon has been found in earlier studies that compared SSrat with the nomination procedures (Maassen et al., 2000). CDnom, as well as SSrat (see Maassen et al., 1996, Tables 1 and 2), is certainly able to detect individuals whose scores truly fit the operational definitions of being neglected or controversial individuals. However, within the CDnom (and consequently also the CDCrat) procedure scores are standardized within classes into *z*-scores along the social impact dimension. In studies where the scores of a group of individuals vaguely meet the definitions, the cutoff limit of 1 *SD* will designate a larger number of individuals as extreme on this dimension. These assignments are doubtful and prove to inflate particularly the size of the neglected category. The classification will prove unstable at a subsequent assessment.

Doubtful assignments to the neglected categories cannot be held responsible for the even greater difference between the stabilities of SSrat and the Newcomb and Bukowski procedure, since the neglected and controversial categories delivered by these two procedures are more similar sized. Lower stability of the popular and rejected categories also decreases the overall stability of this procedure.

In summary, the present study shows that CDCrat and SSrat in particular are useful techniques for the analysis of data collected with the Asher et al. (1979) procedure. The study also specifically shows clear differences between the results of CDCrat and SSrat. Researchers who intend to study neglected children will prefer CDCrat as the classification procedure, since CDCrat assigns considerably larger numbers of individuals to the neglected category. It should be noted that some scholars argue that neglect may not be a meaningful sociometric category (Rubin, Hymel, LeMare, & Rowden, 1989). For researchers who agree with this contention or who are not interested in the neglected or controversial categories, the aspect of stability may become important. SSrat then proves to be a valuable tool, since in this and previous studies it consistently showed a considerably higher stability than did CDnom, CDCrat and the Newcomb and

Bukowski procedure. We have amply argued that the present study provides strong evidence that SSrat is the procedure with the higher overall reliability.

The issues of temporal stability and validity of sociometric status addressed in this study are not only of methodological importance, but also have conceptual or theoretical implications. Recent research for example, is not only aimed at examining the stability of peer status, but also focuses on the antecedents (in child and social environment) and consequences of having a stable peer sociometric status, and specifically, of being constantly rejected (Cillessen et al., 2000). It is evident that the conclusions drawn from these studies depend upon the reliability and validity of the sociometric status classifications that are used. Moreover, several studies have been conducted on the effects of interventions (coaching, social-skills training, participation in peer involvement groups, etc.) on the social behavior and social status of children. It is evident that conclusions concerning sociometric-status change should not be the consequence of instability of sociometric measures. Hence, for future studies it may be interesting to examine the sensitivity of different sociometric-status classification procedures, including the SSrat procedure, in detecting the hypothesized improvements resulting from the intervention before jumping to conclusions.

REFERENCES

Asher, S. R., & Hymel, S. (1981). Children's social competence in peer relations: Sociometric and behavorial assessment. In J. D. Wine & M. D. Smye (Eds.), *Social competence*, 125-157. New York: Guilford.

Asher, S. R., Singleton, L. C., Tinsley, B. R., & Hymel, S. (1979). A reliable sociometric measure for preschool children. *Developmental Psychology*, **15**, 443-444.

Bronfenbrenner, U. (1944). A constant frame of reference for sociometric research. Part II: Experiment and inference, Sociometry, 7, 40-75.

Cillessen, A. H. N., Bukowski, W. M., & Haselager, G. J. T. (2000). Stability of sociometric categories. In A. H. N. Cillessen, & W. M. Bukowski (Eds.): *Recent advances in the measurement of acceptance and rejection in the peer system. New directions for child and adolescent development (nr.88).* San Francisco: Jossey-Bass.

Cillessen, A. H., Van IJzendoorn, H. W., Van Lieshout, C. F., & Hartup, W. W. (1992). Heterogeneity among peer rejected boys: Subtypes and stabilities. *Child Development,* **63**, 893-905.

Cohen, J. (1992). A power primer. *Psychological Bulletin,* **112**, 155-159.

Coie, J. D., & Dodge, K. A. (1983). Continuities and changes in children's social status: A five-year longitudinal study. *Merril-Palmer Quarterly,* **29**, 261-282.

Coie, J. D., Dodge, K. A., & Coppotelli, H. (1982). Dimensions and types of social status: A cross-age perspective. *Developmental Psychology,* **18**, 557-570.

Dunnington, M. J. (1957). Investigation of areas of disagreement in sociometric measurement of preschool children. *Child Development,* **28**, 93-102.

Frederickson, N. L., & Furnham, A. F. (1998). Sociometric classification methods in school peer groups: A comparative investigation. *Journal of Child Psychology and Psychiatry,* **39**, 921-933.

Gronlund, N. E., & Anderson, L. (1957). Personality characteristics of socially accepted, socially neglected and socially rejected junior high school pupils. *Educational Administration and Supervision,* **43**, 329-338.

Hartup, W. W., Glazer, J. A., & Charlesworth, R. (1967). Peer reinforcement and sociometric status. *Child Development,* **38**, 1017-1024.

Howes, C. (1988). Peer interaction of young children. *Monographs of the Society for Research in Child Development,* **53**, 1-94.

Hymel, S., Vaillancourt, T., McDougall, P., & Renshaw, P. D. (2002). Acceptance and rejection in the peer group. In P. Smith & C. Hart (Eds.), *Handbook of Child Social Development.* London: Blackwell Publishers.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159-174.

Maassen, G. H., Akkermans, W., & Van der Linden, J. L. (1996). Two-dimensionsional sociometric status determination with rating scales. *Small Group Research,* **27**, 56-78.

Maassen, G. H., & Landsheer, J. A. (1996). *Manual for SSRAT2.0: A program for two-dimensional sociometric status determination with rating scales.* Utrecht, The Netherlands: Utrecht University FSW/ISOR Publications, Method Series MS-96-2.

Maassen, G. H., & Landsheer, J. A. (1998). SSRAT: The processing of rating scales for the determination of two-dimensional sociometric status. *Behavioral Research Methods Instruments, and Computers,* **30**, 674-679.

Maassen, G. H., Van der Linden, J. L., Goossens, F. A., & Bokhorst, J. (2000). A ratings-based approach to two-dimensional sociometric determination. In A. H. N. Cillessen, & W. M. Bukowski (Eds.): *Recent advances in the measurement of acceptance and rejection in the peer system. New directions for child and adolescent development (nr.88).* San Francisco: Jossey-Bass.

McCandless, B. R., & Marshall, H. R. (1957). A picture sociometric technique for preschool children and its relation to teacher judgments of friendship. *Child Development,* **28**, 139-147.

Moore, S., & Updegraff, R. (1964). Sociometric status of preschool children related to age, sex, nurturance-giving, and dependency. *Child Development,* **35**, 519-524.

Newcomb, A. F., & Bukowski, W. M. (1983). Social impact and social preference as determinants of children's peer group status. *Developmental Psychology,* **19**, 856-867.

Newcomb, A. F., & Bukowski, W. M. (1984). A longitudinal study of the utility of social preference and social impact sociometric classification schemes. *Child Development,* **55**, 1434-1447.

Newcomb, A. F., Bukowski, W. M., & Pattee, L. (1993). Children's peer relations: A meta-analytic review of popular, rejected, neglected, controversial, and average sociometric status. *Psychological Bulletin,* **113**, 99-128.

Oden, S., & Asher, S. R. (1977). Coaching children in social skills for friendship making. *Child Development,* **48**, 495-506.

Olson, S. L., & Lifgren, K. (1988). Concurrent and longitudinal correlates of preschool peer sociometrics: Comparing rating scale and nomination measures. *Journal of Applied Developmental Psychology,* **9**, 409-420.

Rubin, K. H., Bukowski, W. M., & Parker, J. G. (1998). Peer interactions, relationships, and groups. In W. Damon (Series Ed.) & N. Eisenberg (Vol. Ed.), *Handbook of child psychology: Vol. 3. Social, emotional, and personality development* (5th ed.) (pp. 619-700). New York: Wiley.

Rubin, K. H., Hymel, S., LeMare, L., & Rowden, L. (1989). Children experiencing social difficulties: Sociometric neglect reconsidered. *Canadian Journal of Behavioural Science,* **21**, 94-111.

Ten Brink, P. W. M. (1985). *De gegeneraliseerde binomiale verdeling als alternatief voor de sociometrische status berekening volgens het probabiliteitsmodel* The generalized binomial distribution as an alternative for sociometric status determination according to the probability model]. Nijmegen, The Netherlands: Unpublished master's thesis, Psychologisch Laboratorium, Nijmegen University.

Terry, R. (2000). Recent advances in measurement theory and the use of sociometric techniques. In A. H. N. Cillessen, & W. M. Bukowski (Eds.): *Recent advances in the measurement of acceptance and rejection in the peer system. New directions for child and adolescent development (nr.88).* San Francisco: Jossey-Bass.

Terry, R., & Coie, J. D. (1991). A comparison of methods for defining sociometric status among children. *Developmental Psychology,* **27**, 867-880.

Thompson, G. G. & Powell, M. (1951). An investigation of the rating-scale approach to the measurement of social status. *Educational and Psychological Measurement,* **11**, 440-455.

Wiley, D. E., & Wiley, J. A. (1970. The estimation of measurement error in panel data. *American Sociological Review,* **35**, 112-117.

Wu, X., Hart, C. H., Draper, T. W., & Olsen, J. A. (2001). Peer and teacher sociometrics for preschool children: Cross-informant concordance, temporal stability, and reliability. *Merrill-Palmer Quarterly,* **47**, 416-443.