The following manuscript represents the accepted version of the article published as:

Marks, P. E. L., Babcock, B., Cillessen, A. H. N., & Crick, N. R. (2013). The effects of participation rate on the internal reliability of peer nomination measures. Social Development, 22, 609-622.

The Effects of Participation Rate on the Internal Reliability of Peer Nomination Measures

Peter E. L. Marks

Austin College

Ben Babcock

The American Registry of Radiologic Technologists

Antonius H. N. Cillessen

Radboud Universiteit Nijmegen

Nicki R. Crick

University of Minnesota – Twin Cities

Author Note

Peter E. L. Marks, Department of Psychology, Austin College; Ben Babcock, American Registry of Radiologic Technologists (ARRT); Antonius H. N. Cillessen, Behavioural Science Institute, Radboud Universiteit Nijmegen; Nicki R. Crick, Institute of Child Development, University of Minnesota – Twin Cities.

The authors would like to thank Stephen Nydick and Amy Luckner for theoretical contributions to this investigation. The views and discussions presented in this research are not necessarily the official views of ARRT.

Correspondence should be addressed to Peter E. L. Marks, Department of Psychology, Austin College, 900 N. Grand Ave., Sherman, TX 75090. Phone: 310-963-1434. Email: pmarks@austincollege.edu

Abstract

Although low participation rates have historically been considered problematic in peer nomination research, some researchers have recently argued that small proportions of participants can, in fact, provide adequate sociometric data. The current study used a classical measurement perspective to investigate the internal reliability (Cronbach's α) of peer nomination measures of acceptance, popularity, friendship, prosocial behavior, and overt aggression. Data from 642 participants attending 10 schools were resampled at different participation rates ranging from 5% to 100% of the original samples. Results indicated that (a) the association between participation rate and Cronbach's α was curvilinear across schools and variables; (b) collecting more data for a given variable (by using unlimited vs. limited nominations, or two versus one items) was significantly related to higher internal reliability; and (c) certain variables (overt aggression, popularity) were more reliable than others (acceptance, friendship). Implications for future research were discussed.

Key terms: Sociometric Status, Methodology, Peers/Peer Relations, Popularity, Friendship

Collecting data within classroom settings has become increasingly difficult in recent years. Greater IRB demands for active consent procedures, distrust of (or apathy toward) academic research on the part of parents, and the use of aptitude tests to determine school funding (which discourages principals and teachers from using class time for anything other than direct teaching) have made it more difficult to collect data in classrooms or from all students in a given classroom. This has created logistical difficulties for researchers of children and adolescents; more schools and families must be contacted to attain target sample sizes. It also creates a major psychometric issue by limiting participation rates — that is, the proportion of students in a given context (classroom, grade, school) who provide data for the study.

Low response or participation rates are problematic in nearly all research for reasons of external validity. Because there is often no way to know whether nonparticipants are systematically different from participants, results from studies with low participation or response rates must be interpreted with caution. In a self-report study with low participation, data are missing for the nonparticipants only. Low participation is particularly problematic, however, for studies involving peer nominations, where the problems involve both external validity and reliability. In peer nomination studies with low participation, data are missing for *all potential participants*, because nonparticipants would have provided information about each peer in the group. To illustrate: suppose Josie participates in a study of peer aggression, but only half of the peers in her class have received parental consent to participate. If the study involves a self-report of aggression, researchers will be able to collect full information about Josie's aggression level – she can complete the survey even if her peers do not. However, if the study involves peer nominations of aggression, 50% participation implies that the researchers can only collect 50% of the relevant data regarding Josie's level of aggression. Thus, the problem posed by low

participation rates in peer nomination studies is one of reliability – it influences the amount of data being collected on each participant for a given criterion.

Although researchers have historically argued that high participation rates (at least 60-70%, according to Cillessen & Marks, 2011) are necessary to obtain reliable peer nomination data (e.g., Crick & Ladd, 1989; Foster, Bell-Dolan, & Berler, 1986), some researchers have recently suggested that nominations from a small subset of participants may be reliable (Prinstein, 2007; Walcott, Upton, Bolen, & Brown, 2008; Zakriski, Siefer, Sheldrick, Prinstein, Dickstein, & Sameroff, 1999). Unfortunately, very little research has directly investigated the effects of participation rate on the reliability of peer nominations. The current study considered peer nominations from a classical test theory framework (Terry, 2000) to investigate the effect of participation rates on the reliability of peer nomination data.

Conceptualizing Peer Nominations as Traditional Measurement

Some researchers have expressed their skepticism about the measurement reliability of "single item measures" in peer nominations. This skepticism is easy to understand – in questionnaires, single item measures may often be unreliable. However, there are no single-item measures in peer nomination research.[1] A peer nomination item is, from a measurement perspective, a series of binary (true-false or yes-no) "test items." A nominee's score for a given peer nomination question reflects the results of *n* items, where *n* is the number of peer nominators (e.g., participating classmates).

Conceptualizing peer nomination for a criterion as a test with *n* items has several implications. First, in most cases, the items in such a test will have a low proportion of endorsement, or be very "difficult." If we consider a choice for an individual as a keyed response

[1] This is not to say that there are no *validity* issues involved in using a single criterion to measure a given construct in a peer nomination measure. This becomes a matter of whether one's operational definition of a construct accurately reflects that construct; that question is well beyond the bounds of the current paper's focus on reliability.

and a non-choice as a non-keyed response, the number of keyed responses (choices) will be very low in proportion to the number of non-keyed responses (non-choices). This occurs because students almost never nominate 40% or more of their classmates for a particular question. This has implications for reliability, because test questions with a 40% to 60% endorsement rate have the highest variance. This higher variance gives a potential for larger item intercorrelations and, thus, higher reliability. Another implication of considering peer nominations as tests is that they are not "count data" as classically understood. Rather, peer nominations are just like any other test score based on aggregating binary items. It can certainly be argued that the "total score" in the classical testing paradigm could be considered as a type of count data. Total score is a summed composite from a limited observation of behavior (Crocker & Algina, 1986) just like count data are a summed composite, because total score involves counting up the number of keyed responses. There is, however, a qualitative difference between a continuous variable based on test or survey data and classic forms of "count data," particularly when using item response theory frameworks that take into account which particular items have a keyed response.

Finally, and most importantly, the measurement conceptualization of peer nomination questions makes it possible to estimate their reliability using internal consistency statistics. Specifically, it allows for the calculation of Cronbach's α, which indicates the extent to which nominators agreed upon which nominees best fit a given criterion. Cronbach's α is the theoretical mean split-half reliability across all possible split-half combinations of items within a given measure, adjusted for the full-sample *N* (Crocker & Algina, 1986). In the context of peer nominations, Cronbach's α reflects the agreement between any two halves of the nominators

regarding the number of nominations received by any given participant.[2] Researchers will want the nominations provided by half of the nominators to correlate highly with the nominations provided by the other half, because the high correlation indicates that the two groups are in relative agreement.

We are not the first to use a classical measurement perspective to investigate the psychometric properties of peer nominations (see, e.g., Bronfenbrenner, 1943; Terry, 2000), nor are we the first to use Cronbach's α for the internal consistency of a sociometric measure (e.g., Gagné, Bégin, & Talbot, 1993; Gordon, 1969).[3] This is, however the first application of these strategies to examine the effect of participation rates on the reliability of peer nominations.

Previous Work on Participation Rates

Thus far, two published studies have directly investigated the effects of various participation rates on the reliability of peer-derived data. Crick and Ladd (1989) investigated fluctuations of sociometric status groupings (see Coie, Dodge, & Coppotelli, 1982) by recalculating group memberships based on random selections of nominators. They concluded that group classifications were unreliable at lower rates, and recommended that researchers include as many classmates as possible in their work. A decade later, Hamilton, Fuchs, Fuchs, and Roberts (2000) randomly resampled 25%, 50%, and 75% of nominators and recomputed continuous scores from a sociometric rating scale (*not* a peer nomination instrument) of liking/disliking at each resampled proportion, concluding (tentatively) that at least 75% of a classroom should participate to obtain reliable data.

[2] There are certainly a wide variety of alternatives to Cronbach's α (see Green, 2003 and Sijtsma, 2009 for example), but the reliability estimate itself is not the focus of the current work.

[3] Modern sociometric researchers often use Cronbach's α when comparing multiple items regarding a given criterion (e.g., using 4 nomination items as a measure of overt aggression). In this case, however, we are referring to the use of α to look at the internal consistency *among nominators* within *either* single- *or* multiple-item nomination measures.

The conclusions from these studies contrast with more recent arguments and findings suggesting that low participation rates can be reliably used in peer nomination research. However, several of these recent justifications of low participation rates are based on findings and claims within unpublished manuscripts (e.g., Angold et al., 1990, ctd. in Prinstein, 2007; Nukulkij et al., 2002, ctd. in Sandstrom & Cillessen, 2003; Terry et al., 1998, ctd. in Zakriski et al., 1999). As far as we know, only one published study has directly concluded, based on empirical findings, that peer nomination data derived from low participation rates can be reliable. Prinstein (2007) demonstrated that nomination counts from a full sample of adolescents would correlate moderately to highly with nomination counts derived from subsamples of 10% of the adolescents. Based on these correlations, Prinstein (2007) concluded that small subsamples could provide reliable nomination data. However, because the subsample participants were also included in the full sample, the correlations may have overestimated the true correspondence, due to item overlap correlation (see Guilford, 1936; Hsu, 1992). In other words, the exact correspondence between sociometric data from small subsamples and a full sample may not be known yet.

The Current Study

The current study investigated the effects of different participation rates on the internal consistency reliability of peer nominations. Building on the methods of Crick and Ladd (1989) and Hamilton et al. (2000), data from 5th grade students from 10 schools were repeatedly resampled to simulate different participation rates, ranging from 5% to 95%. Cronbach's α was the primary measure of reliability calculated for each participation rate in each school.

Five variables were measured by peer nominations: peer acceptance, friendship, popularity, overt aggression, and prosocial behavior. These variables are commonly assessed in

the developmental literature using peer nominations. Acceptance and friendship are affective judgments, whereas popularity is based on an individual's social reputation in the peer group (Cillessen & Marks, 2011). These variables provide an ideal contrast to overt aggression and prosocial behavior, which are based on distinct and observable behaviors.

Although acceptance and friendship require nominators to make an individual judgment about their affective reactions toward others, these variables are commonly used by researchers as continuously distributed latent variables. Many sociometric studies correlate acceptance and friendship with social behaviors and other personal characteristics. When acceptance and friendship are used in this way, Cronbach's α is an appropriate measure of reliability, because they are not used to indicate individual relationships, but the overall amount that each participant is liked by classmates. Obviously, when friendship nominations are used to study dyadic relationships (as opposed to group processes), reliability is irrelevant.

Hypotheses

The primary hypothesis of this study was that lower participation rates would result in lower internal reliabilities. We did not expect a particular cutoff at which measures will be unreliable, but we did expect that very low participation rates (below 20%) would provide unreliable data. In addition to this main hypothesis, we examined several other hypotheses:

- Unlimited nominations would lead to more reliable data than using limited nominations. Friendship was assessed with *both* limited *and* unlimited nominations. We expected higher reliability, across participation rates, for unlimited nominations (see Alain & Bégin, 1987).
- Two questions for a given criterion would yield more reliable results than one question. We used two peer nomination questions for prosocial behavior to test this hypothesis.

- Reliability would be higher for specific behaviors than for affective reactions, because the nomination questions are requesting qualitatively different information (Babad, 2001). Acceptance and friendship nominations request information about individual feelings or relationships to peers. Overt aggression and prosocial behavior nominations, in contrast, request information about concrete, observable tendencies of peers. We expected higher participant agreement and reliability for behavior nominations than for affective nominations (Babad, 2001).
- Nominations of popularity would be more reliable than nominations of acceptance or friendship. Popularity is based on the group consensus of a peer's reputation (Cillessen & Marks, 2011). Nomination scores should reflect this consensus in high inter-rater agreement. This would support previous literature indicating higher test-retest reliability and stability for popularity than for affective nominations (see Cillessen & Marks, 2011, for a review), and would parallel findings of Prinstein (2007).

Method

Participants

Participants were 642 fifth graders who were enrolled in a larger longitudinal study of peer relationships. The participants (i.e., potential nominees) formed the entire fifth grade populations at 10 elementary schools, each ranging in size from 38 to 90 students. Completion rates for the peer nomination instrument ranged from 82% to 97% (M = 91%) across schools. All nominators completed measures in their own classrooms during the spring of their fifth grade year. Participants were 47.9% female and predominantly Caucasian (74.0%), African American (15.0%), and Hispanic (9.1%).

The elementary schools from which the data was collected included all elementary

schools in a mid-sized New England town with an overall population of approximately 50,000. The town included mostly lower-middle-class and middle-class neighborhoods.

Peer Nomination Measure

The peer nomination procedure used focused, single-item nominations (see Becker & Luthar, 2007; Parkhurst & Asher, 1992) to assess peer views of affective, reputational, and behavioral characteristics of peers. Items relevant to the current investigation included nominations of friendship ("the people who are your best friends"), liking/peer acceptance ("the people in your grade you like the most"), popularity ("the people in your grade who are the most popular"), overt aggression ("the people in your grades who start fights, say mean things, and/or tease others"), and prosocial behavior (2 items, "the people in your grade who cooperate, share, and help others," and "the people in your grade who are leaders and good to have in charge;" r = .88). In addition to these items, nominators were asked to nominate peers on the basis of disliking/rejection, social isolation, general victimization, relational aggression (2 items; r = .51), relational victimization, relational inclusion, and unpopularity. For each item, nominators were given a roster of all same-grade peers within their school from which to make nominations, and were asked to circle the code numbers of all peers who fit the item. Nominators were permitted to nominate an unlimited number of peers for each item; self-nominations were removed prior to calculation of scores. For the friendship item only, nominators were additionally asked to indicate their top 5 friends, in order of preference.

Data Preparation

As in previous research with peer nominations, raw scores for each item were calculated based on the sum of the nominations received by each participant. For friendship, two raw scores were calculated: the overall number of friendship nominations received based on unlimited

nominations, and a "top 3" friends nomination based on the number of times a participant was named as being a peer's first, second, or third best friend. Limited friendship was based on 3 nominations (as opposed to 5), because this is the most common number of nominations allowed for a limited nomination procedure (Cillessen & Marks, 2011).

Two raw scores were also calculated for prosocial behavior. The "one-item" score represented the total number of nominations received for the "cooperate, share, and help others" item. The "two-item" score was based on the sum of the total number of nominations for both prosocial behavior items (see above).

The mean number of nominations received for each item was similar across the 10 schools for each of the main study variables: liking (M = 9.21, SD = 1.12), popularity (M = 8.00, SD = 1.50), overt aggression (M = 8.15, SD = 1.47), prosocial behavior (M = 11.08, SD = 1.86 with one item, M = 20.21, SD = 3.71 with two items), unlimited friendship (M = 9.46, SD = 1.96), and "top 3" friends (M = 2.77, SD = 0.10).

The analyses were guided by a classical test theory framework. The nomination data for each sociometric question in each school were converted to a binary matrix with columns designating nominators and rows designating nominees. This was the "persons by items" matrix generally analyzed in psychometric procedures. A 1 in a cell indicated that the nominee (row) was named the nominator (column) for the question; a 0 indicated that the nominee was not named by that nominator. This organization of the data treated each nominee as a participant taking an exam with difficult items (i.e., having a relatively low chance of "getting an item correct," or, in this case, being nominated for that question). Thus, the number of nominees (same-grade students in each school) was considered to be the sample size N. The number of nominators was n, the number of test items on which each nominee had a score in the matrix for

a given sociometric question.

In addition to allowing an efficient resampling of nomination data for different groups of nominators (see below), this preparation of the data made it possible to calculate Cronbach's α internal consistency coefficients. An alternative reliability measure, Guttman's λ_2 (Sijtsma, 2009), was also used for comparison, and yielded results very similar to Cronbach's α. The authors also analyzed the smaller samples using the Spearman-Brown equation (Brown, 1910; Spearman, 1910) and again found similar results as with Cronbach's α. The fact that the reliability estimates from the smaller samples predicted the full-sample reliability with some accuracy provides evidence that Cronbach's α functioned properly with the current data sets.

Resampling Procedure

Following the terminology of Hamilton et al. (2000), the term "standard level" was used to refer to the full-sample participation rate (ratio of number of nominators to number of nominees) in the collected data. In other words, if a given school included 95% actual participation, we considered that participation rate to be 100% for the purpose of the resampling procedure. In order to simulate lower rates of participation, a proportion of nominators from the full data sample, say .80, was randomly chosen without replacement. This resampled dataset had 80% of the nominators from the original sample. The participation rate for this example was, thus, 20% less than the standard level. The resampling procedure intentionally excluded 20% of the nominators in order to simulate data conditions for which the participation rate was lower. After selecting the sample, we recalculated total scores and Cronbach's α. We used sampling proportions from .05 to .95 in intervals of .05 (19 levels in all). In order to obtain stable estimates, we conducted 1,000 replications for each of the 19 sampling rates for each of the 10 schools (190,000 replications in total).

Results

Preliminary Analysis: Differences Between Participants and Nonparticipants

Given that all 5th grade students were included as nominees in the peer nominations collected from each of the 10 schools, it was possible to compare participants who completed nomination measures (nominators) to those who did not (non-nominators), in terms of sociometric scores received for every sociometric item. If scores did not differ between them, it would be appropriate to randomly exclude nominators during the resampling procedure using the assumption of missing completely at random (MCAR, Schafer & Graham, 2002). If scores were significantly different between nominators and non-nominators, however, then a missing at random (MAR) framework should be used, with some nominators having a greater chance of being excluded depending on their sociometric scores.

Independent samples *t*-tests were run to compare nominators and non-nominators on each sociometric variable. Given that the majority of participants completed measures in any given school, nominators and non-nominators were analyzed across schools (nominator $N = 580$, non-nominator $N = 61$). In addition to the items of our central interest (friendship, liking, popularity, prosocial behavior, overt aggression), nominators and non-nominators were compared on rejection, social isolation, general victimization, relational victimization, relational aggression, relational inclusion, and unpopularity. Given the large number of *t*-tests being performed, a significance level of $p < .01$ was used.

Nominators and non-nominators did not differ significantly on any of the sociometric variables (t ranged from -1.98 to 1.54, p ranged from .05 to .78). Thus, the resampling procedure in all of the following analyses was conducted assuming that participants were MCAR.

Alpha Levels

Following the data preparation and resampling procedures described above, a base Cronbach's α for each item was determined for the standard level of participation in each school (i.e., including all nominators who actually participated from each school). Mean Cronbach's α's for each question were calculated for each resampled participation rate in each school. That is, the mean Cronbach's α's for each question were the mean of 1,000 replications for each participation rate in each school.

Graphs comparing participation rates to mean α levels in each school were created for each variable and are presented in Figure 1, which presents the minimum, maximum, and median alphas across schools for each variable. The graphs clearly show a curvilinear association between participation rate and α. The reliability of the sociometric items increased rapidly as the resampled participation rate went from 0 to .50. Reliability continued to increase as the participation rate increased above .50, but the gains were not as large.

A few trends specific to individual items or schools are worth noting. First, some reliability lines were shorter than others at the high end of the graphs. This occurred because different schools had differing maximum participation rates. The differing maximum participation rates do not affect the message of this research, however, because the reliability trends held true across all schools despite their actual empirical participation rates. Second, the reliabilities for the "Top 3 Friends" item were quite low and did not increase quickly in the low ranges of participation rate. By limiting nominations to three friends at most, the "Top 3 Friends" variable restricted the range of possible scores. When viewing raters as items on an exam, limiting nominations makes each exam item extremely difficult. Very difficult items have low variances, which can restrict the reliability coefficients (Crocker & Algina, 1986). Third, reliability for prosocial behavior was consistently higher, though not as dramatically, when it

was based on two items rather than one. Finally, the reliabilities for overt aggression, prosocial behavior, and popularity were higher overall than the reliabilities for liking and friendship. Reliability also increased faster in the variables measuring behavior and reputation than in those measuring affective reactions.

Interval Band Analysis

In order to evaluate the effectiveness of using more data (unlimited vs. limited friendship nominations and two vs. one prosocial behavior questions), we constructed ± 1 standard deviation bands around the mean reliabilities. To do this, we found the standard deviation of the resampling replications for each participation rate. We then found the mean minus the standard deviation for the higher reliability condition and the mean plus the standard deviation for the lower reliability condition. Distributions whose standard deviation bands do not overlap will have an extraordinarily large effect size difference (in this study, between 1.75 and 2.00) according to Cohen's (1988) guidelines. For friendship nominations, the standard deviation bands did not overlap for any condition where the participation rate was greater than .65 (range .30 to .65). The prosocial behavior standard deviations never overlapped for any condition with a participation rate greater than .75 (range .50 to .75). Conditions that gathered more data were clearly more reliable than conditions that gathered less.

Discussion

The results of this study supported the hypothesis that higher participation rates are associated with higher reliabilities of peer nomination data. For each school and variable, Cronbach's α was higher when participation was higher. In all cases, the association between participation rate and α was curvilinear, with the steepest curve at lower participation rates.

Also as expected, nominations of popularity were more reliable than nominations of

liking (acceptance). Contrary to expectations, however, *some* variables showed acceptable reliability (above .80) at *some* schools with a participation rate as low as 10%-20%. These results supported findings by Prinstein (2007) and others that a small subsample of nominators may provide reliable peer nomination data, *at least for certain variables in certain contexts*. It was also the case, however, that data were not reliable in the 10%-20% range for some schools even for overt aggression, the most reliable variable in the study. Given that a higher participation rate *always* resulted in more reliable data, experimenters should not purposely gather little data.

The results supported all other hypotheses. As expected, collecting more data for a given criterion increased reliability. This was especially true for friendship, where we compared unlimited nominations with limited "top 3" nominations. Friendship nominations were dramatically more reliable with the unlimited procedure. Similarly, although less striking, prosocial behavior was significantly more reliable when measured with two peer nominations rather than one, particularly with participation rates of 60% or greater.

A higher number of nominations increased reliability only for certain criteria and not in an absolute sense. For example, unlimited friendship nominations were more reliable than "top 3" friend nominations, but still less reliable than overt aggression nominations despite the fact that participants received on average more unlimited friendship nominations (9.46) than overt aggression nominations (8.15). In fact, certain variables were consistently more reliable than other variables. As hypothesized, popularity, overt aggression, and prosocial behavior tended to be more reliable than liking and friendship. This may reflect the fact that liking and friendship are based on individual affective reactions, whereas aggression and prosocial behavior are based on observable, concrete behaviors (Babad, 2001), and popularity is based on the agreed-upon consensus of the peer group (Cillessen & Marks, 2011).

Many methodologists (the authors included) believe that the measurement process entails quantifying an unobservable construct through observable behaviors specified by an operational definition (Cronbach & Meehl, 1955). It is possible that "the extent to which one has more friends in the classroom" may not be a reliable operational definition for friendship when aggregated across friend groups within a classroom or grade. Unlike nominations for more easily operationalized traits like overt aggression, whether two peers are friends depends on their behaviors towards each other as well as their interpretation of these behaviors. The way methodologists currently operationally define friendship in instructions may not be strong enough to account for all ipsative factors involved. An alternative explanation is that students in a classroom rarely fully agree on who their friends are. If this is the case, one may never be able to achieve extremely reliable measurement as long as reliability is cast within the mold of inter-rater agreement, as it is with a Cronbach's α type of measure.

Limitations

There were some limitations to this study, most of them related to the sample in which the data were gathered. The data came from a single age group in 10 schools in a single city. Similarly, the data were based on a limited range of grade sizes (38 to 90). Future research will need to investigate whether our results generalize to other age groups, other locations, or in very small classrooms or very large grades.

In the current sample, there were no differences between participants and nonparticipants in terms of numerous nomination scores, allowing us to exclude nominators completely at random when simulating missing data. However, systematic differences between participants and nonparticipants may be observed in other studies peer nomination studies. Such differences may affect the reliability and validity of peer nomination data. Although it was appropriate to use an

MCAR framework in this study, the results reflected random, and not systematic, nonparticipation. MAR or MNAR missing data could have different effects on reliability. It will be interesting to examine this type of peer nomination measurement from an item response theory (IRT) perspective, which often treats missing data as MAR.

Finally, it should be emphasized that any results or conclusions regarding acceptance and friendship nominations reflect the use of these nominations to calculate continuous latent variables. Reliability would not be an appropriate consideration, for example, when using friendship nominations to identify specific dyadic relationships. However, these results suggest that reliability may be an issue when determining how well-liked a student is, or how many friendships an adolescent has, within the classroom.

Recommending "Cutoffs"

It is tempting to want to provide certain "cutoffs," that is, recommendations for minimum appropriate values for reliability, participation rate, or other factors. This is a fine line to walk. On the one hand, almost no methodological data allows for a clear-cut value for which researchers should strive; our data may tell us how to measure reliability, but it provides no indication of what constitutes "good" versus "bad" reliability. On the other hand, if no guidelines or recommendations are provided, future researchers may use the current study to justify inappropriate methodological decisions; for example, they may say that their 20% participation rate provides reliable overt aggression nomination data simply because our results showed that a 20% participation rate *could* result in reliable data (see Walcott et al., 2008).

As such, we feel it is appropriate to provide certain guidelines for evaluating reliability in future peer nomination research. However, it should be noted that these recommendations are based on both methodological expertise and the results of the data presented in this paper.

Given that we used a traditional test theory framework for the reliability of peer nominations, it seems reasonable to recommend cutoffs generally used in other survey studies of human behavior; that is, accepting a minimum Cronbach's α of .60 and considering values of .80 or greater to indicate satisfactory reliability. We should keep in mind, however, the interpretation of reliability estimates. If a dataset has an estimated reliability of .60, the measurement of this group consists of 60% "signal" and 40% "noise," or error variance. This error will greatly limit the effectiveness of any number of statistical procedures, including the ability to detect group differences and limiting the size of correlation coefficients (Crocker & Algina, 1986). We assume that most researchers will want to exceed this "minimum" reliability level.

Using these criteria, we conclude that it is impossible to recommend a context-free "cutoff" or "target" participation rate. The generally accepted minimum rate of 60%-70% (Cillessen & Marks, 2011) appears, from a reliability perspective, to be unsupported by the current study. Certainly, overt aggression and popularity showed good reliabilities with participation rates as low as 40% in all schools, but for friendship, some schools never achieved a reliability of .60, even with participation above 85%. There was considerable variability in the maximum reliability across measures, indicating that participation rate alone provides a limited indication for the reliability of a peer nomination measure.

Overall, then, we recommend that studies using peer nomination measures should report reliability for these measures, particularly when participation rates are low or moderate. Sometimes logistical constraints make it difficult to obtain data from a high proportion of a classroom or grade. In these cases, calculating and reporting a high α will support the data.

Collecting as Much Data as Possible to Increase Reliability

The current study demonstrated that the numbers of peer nominations can affect their

reliability. We recommend that future peer nomination studies use (a) an unlimited nomination procedure, and/or (b) multiple nomination questions for each criterion. Concerning (b), it is possible that combining the results of more than one question could lead to lower reliability if the first question is very good and the additional ones do not correlate well with the first one. Adding more items into the mix, however, will more likely improve reliability. Additional strategies of increasing nominations may be helpful. For example, providing nominators with a full list of peers for each item and asking them to *circle* names or code numbers is easier and less fatiguing than matching names to ID numbers and reporting these numbers for each item.

Allowing grade-wide nominations (as opposed to class-wide) is also an appropriate way to collect more data. In our data a larger sample size (more nominators and nominees) was related to higher reliability of the nominations. This indicates that, when possible, grade-wide nominations are generally preferable to within-classroom nominations. Grade-wide nominations are likely to be more appropriate at older ages and particular cultural contexts. Adolescents in the US, for example, are likely to know peers from throughout the grade (e.g., Cillessen & Mayeux, 2004) but this is not necessarily the case in US elementary schools or in secondary schools in countries outside North America (e.g., de Bruyn & Cillessen, 2006).

The overarching measurement concept that unites these issues is: having responses to more test items tends to give higher quality measurement. Whether a researcher accomplishes this by combining multiple items or by obtaining more raters, having more responses nearly always leads to better measurement.

Future Directions

The most relevant future need is for researchers to report reliability analyses for their peer nomination data. This information will help to support, clarify, or contradict the current results. It

is vital that peer relationships researchers establish a base of information regarding practices that will maximize reliability across different samples, age groups, locations, school systems, countries, and cultural contexts.

Future studies should also examine the effects of systematic missingness (MAR or MNAR) on the reliability of sociometric data. The type of missingness could influence reliability. For example, if certain types of children are systematically less likely to provide peer nominations, reliability estimates may be artificially high due to the homogeneity of nominators. However, regardless of the effects of different types of missingness, we expect that the general trend found in this study (more data means higher reliability) will still hold true.

Future simulation work should investigate which reliability estimation method is best. As mentioned, there are various alternatives to Cronbach's α. It is possible that one of them is better for certain peer nomination situations. Future simulation and real data research with larger samples could also examine the more nuanced question of whether reliability of peer nominations is driven by participation rate or the absolute number of nominators. This study focused on sample sizes less than 100. It was difficult to tell whether the absolute number of nominators or the participation rate drove the changes in reliability. Participation rates may not be as strong when dealing with a much larger pool of potential nominators.

Finally, future research should contrast internal reliability versus test-retest reliability of peer nominations. There are advantages and disadvantages to each approach. For example, internal reliability may be biased by low samples sizes (Gordon, 1969), whereas test-retest reliability may be inflated by memory effects. Because internal reliability can be calculated without an additional administration, it is logistically more feasible in many studies, especially if they are not longitudinal. Regardless of the measurement and logistical considerations, however,

it is interesting to note that Jiang and Cillessen's (2005) meta-analysis reported an average test-retest reliability of .72 for acceptance/liking nomination measures – a coefficient that is in line with the current study's internal reliability findings for samples with high participation rates (see Figure 1). Perhaps both measures of reliability are appropriate for this type of data.[4]

[4] On the other hand, test-retest reliability might be a problematic measure of reliability when the participation rate is low or when requesting limited nominations, given that zero-inflation (i.e., range restriction) can cause problems with test-retest correlations.

References

Alain, M., & Bégin, G. (1987). Improving reliability of peer-nomination with young children. *Perceptual and Motor Skills, 64*, 1263-1273.

Babad, E. (2001). On the conception and measurement of popularity: More facts and some straight conclusions. *Social Psychology of Education, 5*(1), 3-29.

Becker, B. E., & Luthar, S. S. (2007). Peer-perceived admiration and social preference: Contextual correlates of positive peer regard among suburban and urban adolescents. *Journal of Research on Adolescence, 17*(1), 117-144.

Bronfenbrenner, U. (1943). A constant frame of reference for sociometric research. *Sociometry, 6*(4), 363-397.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*, 296-322.

Cillessen, A. H. N., & Marks, P. E. L. (2011). Conceptualizing and measuring popularity. In A. H. N. Cillessen, D. Schwartz, & L. Mayeux (Eds.), *Popularity in the peer system* (pp. 25-56). New York: Guilford Press.

Cillessen, A. H. N., & Mayeux, L. (2004). From censure to reinforcement: Developmental changes in the association between aggression and social status. *Child Development, 75*, 147-163.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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

Crick, N. R., & Ladd, G. W. (1989). Nominator attrition: Does it affect the accuracy of

children's sociometric classifications? *Merrill-Palmer Quarterly, 35*, 197-207.

Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory*. Belmont, CA: Wadsworth.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

de Bruyn, E. H., & Cillessen, A. H. N. (2006). Popularity in early adolescence: Prosocial and antisocial subtypes. *Journal of Adolescent Research, 21*(6), 607-627.

Foster, S. L., Bell-Dolan, D., & Berler, E. S. (1986). Methodological issues in the use of sociometrics for selecting children for social skills research and training. In R. J. Prinz (Ed.), *Advances in behavioral assessment of children and families* (Vol. 2, pp. 227-248). Greenwich, CT: JAI Press.

Gagné, F., Bégin, J., & Talbot, L. (1993). How well do peers agree among themselves when nominating the gifted or talented? *Gifted Child Quarterly, 37*, 39-45.

Gordon, L. V. (1969). Estimating the reliability of peer ratings. *Educational and Psychological Measurement, 29*, 305-313.

Green, S. B. (2003). A coefficient alpha for test-retest data. *Psychological Methods, 8*, 88-101.

Guilford, J. P. (1936). *Psychometric methods.* New York: McGraw-Hill.

Hamilton, C., Fuchs, D., Fuchs, L. S., & Roberts, H. (2000). Rates of classroom participation and the validity of sociometry. *School Psychology Review, 29*, 251-266.

Hollander, E. P. (1957). The reliability of peer nominations under various conditions of administration. *Journal of Applied Psychology, 41*, 85-90.

Hsu, L. M. (1992). Correcting correlations of personality scales for spurious effects of shared items. *Multivariate Behavioral Research, 27*, 31-41.

Hsu, L. M. (1994). Item overlap correlations: Definitions, interpretations, and implications. *Multivariate Behavioral Research, 29,* 127-140.

Jiang, X. L., & Cillessen, A. H. N. (2005). Stability of continuous measures of sociometric status: A meta-analysis. *Developmental Review, 25*, 1-25.

Parkhurst, J. T., & Asher, S. R. (1992). Peer rejection in middle school: Subgroup differences in behavior, loneliness, and interpersonal concerns. *Developmental Psychology, 28*(2), 231-241.

Prinstein, M. J. (2007). Assessment of adolescents' preference- and reputation-based peer status using sociometric experts. *Merrill-Palmer Quarterly, 53*, 243-261.

Sandstrom, M. J., & Cillessen, A. H. N. (2003). Sociometric status and children's peer experiences: Use of the daily diary method. *Merrill-Palmer Quarterly, 49*, 427-452.

Schafer, J. S., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147-177.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107-120.

Spearman, C. C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271-295.

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 acceptence and rejection in the peer system. New Directions for Child and Adolescent Development* (Vol. 88, pp. 27-53). San Francisco: Jossey-Bass.

Walcott, C. M., Upton, A., Bolen, L. M., & Brown, M. B. (2008). Associations between peer-perceived status and aggression in young adolescents. *Psychology in the Schools, 45*,

550-561.

Zakriski, A. L., Seifer, R., Sheldrick, R. C., Prinstein, M. J., Dickstein, S., & Sameroff, A. J. (1999). Child-focused versus school-focused sociometrics: A challenge for the applied researcher. *Journal of Applied Developmental Psychology, 20,* 481-499.

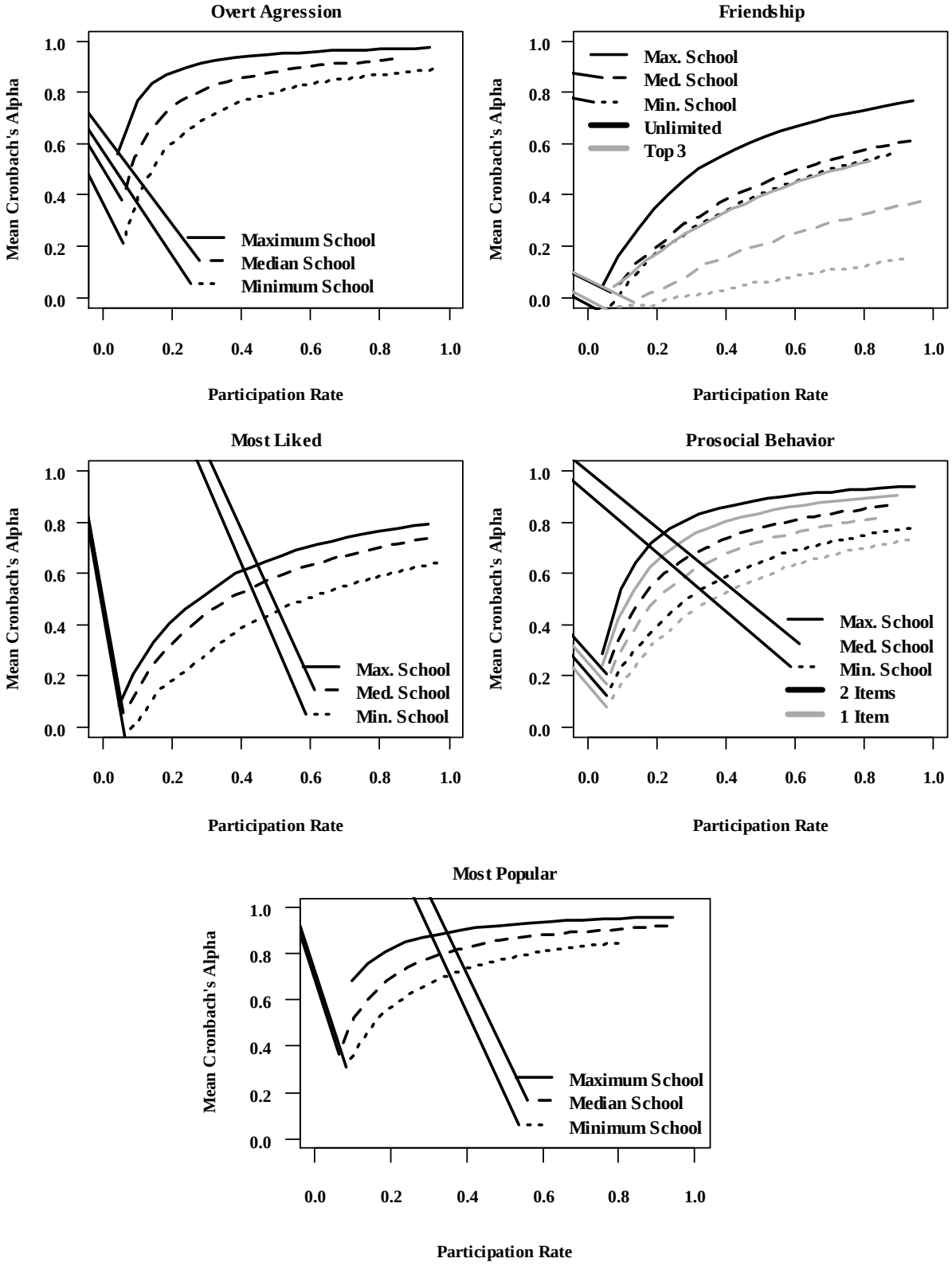

Figure 1. Mean Cronbach's alpha statistics for the lower-bound, upper-bound, and median schools.