Limited Nomination Reliability Using Single- and Multiple-Item Measures

Ben Babcock

The American Registry of Radiologic Technologists

Peter E. L. Marks

Austin College

Nicki R. Crick

University of Minnesota – Twin Cities

Antonius H. N. Cillessen

Radboud Universiteit Nijmegen

Author Note

Ben Babcock, The American Registry of Radiologic Technologists (ARRT); Peter E. L. Marks, Department of Psychology, Austin College; 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 Peter Ralston for theoretical contributions to this investigation. The views and discussions presented in this research are not necessarily the official views of ARRT.

Address correspondence to Ben Babcock, ARRT, 1255 Northland Drive, St. Paul, MN 55120. Email: ben.babcock@arrt.org

We dedicate this research to the memory of Nicki Crick, whose research contributions greatly strengthened the field of human development.

Abstract

This paper examines a variety of reliability issues as related to limited nomination sociometric measures. Peer nomination data were collected from 77 6th grade classrooms. Results showed that, although some single-item peer nomination measures were relatively reliable, many single-item peer nomination measures using limited nominations were quite unreliable. Overt aggression nomination items were the only set of single-item measures where mean classroom reliability estimates were .75 or greater. Combining multiple items led to substantially better reliability, as combining the two least reliable items for a category into a single measure made the composite more reliable than the most reliable single measure. Having more nominators in the sample also increased reliability. The limited nomination items overall tended to be less reliable than similar unlimited nomination items from other studies. The authors end with recommendations for obtaining the most reliable peer nomination data possible from a study.

Keywords: Peer Nominations, Reliability, Aggression, Prosocial Behavior

Limited Nomination Reliability Using Single- and Multiple-Item Measures

Researchers have used peer nomination methods to assess and study children and adolescents since the early twentieth century. The “Guess Who” behavioral assessments of Hartshorne and May (1929) and the peer affiliation (“sociometric”) research of J. L. Moreno (1934) popularized peer nominations. For typical peer nomination procedures, participants are given a list of peers and asked to note which peers fit a criterion (e.g., “Which person do you like most?”). Today, peer nomination measures are commonly used in social development research to assess a variety of constructs (Babad, 2001). However, the specific methodologies of peer nomination studies may differ, and simple decisions (question wording, score calculation, etc.) can have important effects on the collected nomination data and their analysis (Terry, 2000).

One methodological decision made by researchers collecting peer nomination data is whether to allow limited or unlimited nominations—that is, whether to specify the number of peers (traditionally three; Terry, 2000) that a participant may choose for a given criterion or to allow the participant to identify any number of peers. Historically, limited nominations were more common due to the fact that results using unlimited nominations were difficult to analyze. Classic work on the measurement of social status (see Bronfenbenner, 1944; Zeleny, 1940) and sociometric status groupings (Coie, Dodge, & Coppotelli, 1982; Peery, 1979) utilized limited nominations, in addition to more recent studies investigating such topics as peer admiration (Becker & Luthar, 2007) and the difference between overt aggression and relational aggression (Crick & Grotpeter, 1995). Today, unlimited nominations are increasingly preferred by sociometric researchers (e.g., Malcolm, Wensen-Campbell, Rex-Lear, & Waldrip, 2006; Waldrip, Malcolm, & Jensen-Campbell, 2008; Wang, Houshyar, & Prinstein, 2006), based on the consensus that they are more ecologically valid and allow for a more normal distribution of nomination data (Cillessen & Marks, 2011; Parkhurst & Asher, 1992). A small number of studies support this consensus by indicating that unlimited nomination scores may be more reliable than limited nomination scores. Eng and French (1948) compared results from unlimited and limited nominations of roommate preferences with results from paired comparisons and ranking methods, which (though lengthy) are optimal sociometric methods from a measurement perspective (see Terry, 2000). The researchers found that results of paired comparisons and rankings were more highly correlated with unlimited than with limited nomination data. Although interesting, these results should be interpreted with caution, given that Eng and French's sample was small (N = 32) and that the results provided only an indirect measure of reliability.

More recently, Marks, Babcock, Cillessen, and Crick (2013) investigated the internal reliability of sociometric measures using Cronbach's alpha. They noted that unlimited friendship nominations resulted in significantly more reliable data than limited friendship nominations. Limited friendship nominations did not reach acceptable levels of reliability in any of the schools studied (higher than .70 or .80 for certain research purposes, though acceptable reliability in the peer nomination context is certainly debatable; Lattin, Carroll, & Green, 2003, Ch. 6; Furr & Bacharach, 2008, Ch. 5). However, there are two issues that limited the generalizability of these results. First, friendship was the only variable assessed with both limited and unlimited nominations (all other variables were assessed with unlimited nominations only), and friendship was the least reliable criterion, even *with* unlimited nominations. It is unknown whether a more robust variable (such as overt aggression, which Marks et al. showed with unlimited nominations to be highly reliable) might have been sufficiently reliable with limited nominations. Second, many modern limited nomination procedures combine multiple nomination items. For example, although *friendship* is generally assessed with one item (e.g., “Who are your best friends?”), *overt aggression* is often assessed with multiple items capturing different aspects of the behavior. One item might ask about peers who “hit,” while another might ask about pushing behavior (e.g., Rose, Swenson, & Waller, 2004).

The aim of the current study was to follow up on Marks et al. (2013) by investigating the reliability of limited peer nomination data, and by determining whether the use of multiple nomination items for a single criterion will increase a criterion’s reliability.

Limited Nominations and Reliability

The theoretical underpinnings for why limited nominations would yield low-reliability measurements make sense when viewed through the analogy of achievement-type tests. If children in a class of 31 could name only three peers per item (say, “Who do you like most?”), the maximum proportion of the class a student could nominate is .10, or 10%. If an achievement-type test question was answered correctly by only 10% of students, this question would be quite difficult. The scores of exams using only a few very difficult items have low variability and, thus, tend to have low reliability (McDonald, 1999).

One potential solution to this reliability problem is to increase the number of “items.” Even if all items on an achievement-type exam are difficult, an exam with 100 to 200 items will produce scores with sufficient variability to yield reasonably reliable results, assuming positive item intercorrelation (Crocker & Algina, 1986). In the case of peer nominations, each nominator can be seen as an “item” from a measurement perspective. Although one cannot simply enroll new nominators in a classroom, one can combine multiple limited peer nomination items into a composite in order to increase the number of unique nominator-item events. The reliability of the measurements will likely increase because of the added data even though the individual raters are "difficult" items (see Marks et al., 2013).

Reliability Estimation and Cronbach's Alpha

Having reliable data is a requirement for good research. Unreliable measurements contain a large amount of measurement error, which limits the size of the relations between variables (Spearman, 1907). Measurement error can have an effect on any test of statistical significance (Liu & Salvendy, 2009); any statistical conclusion drawn from unreliable data would be called into question by the knowledgeable methodologist.

There are several methods available for reliability estimation. Some methods allow for reliability estimation using only one sample of data. The current study used Cronbach's alpha, which is equivalent to the Kuder-Richardson 20 method when data are dichotomous (Furr & Bacharach, 2008). This study used alpha for two main reasons. First, numerous studies in a variety of fields have used alpha to estimate reliability over several decades. Second, Marks et al.'s (2013) recent investigation of other reliability methods with peer nominations found results that were quite similar to the results using alpha.

The equation for the lower-bound reliability estimate using Cronbach's alpha is

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n} \mathrm{var}_i}{\mathrm{var}_{total}}\right), \qquad (1)$$

where n is the number of items (nominators) and var is the variance of a variable (Cronbach, 1951). This reliability calculation begins with perfect reliability ("1") and subtracts the proportion of variance accounted for by looking at individual items in isolation ("Σ var" term). For peer nominations, reliability in the alpha context represents the correlation or synergy between the raters, scaled up appropriately to the number of raters. When ratings are positively correlated, the total variance will be larger than the sum of the individual item variances. Alpha is 0 if the ratings are uncorrelated. The sum of item variances will be larger than the total variance if the correlation between ratings is negative. This causes alpha to be negative, indicating that either Cronbach's alpha is not a good reliability estimate for the data under consideration, that the data are quite unreliable, or both.

Values for "good" and "acceptable" reliability always depend on the context of the data, the strength of conclusions one wishes to draw, and the stakes of any decisions made based on the data. In the current paper, we considered alpha values of .60 to indicate minimum "acceptable" reliability, and alpha values above .80 to indicate "good" reliability. We acknowledge that .60 is lower than levels of acceptable reliability suggested in many places in the literature (e.g., Lattin, Carroll, & Green, 2003, Ch. 6; Furr & Bacharach, 2008, Ch. 5); the lower criterion here accommodates the fact that past studies have shown limited peer nominations to be relatively unreliable (Marks et al., 2013). A reliability level of .60 is not good but is also not terrible in the context of certain limited nomination items. Researchers should certainly strive for reliability levels well above .60.

Reliability for Single-Item Versus Multiple-Item Measures

For peer nomination items, one may calculate alpha by converting nominations into a data matrix where nominators are in the columns and nominees are in the rows. This setup treats nominators as items are treated traditionally in measurement. Cells corresponding to nominations have a value of 1, and all other cells have a 0. Because each nominator is treated as an "item" in the traditional measurement sense, one may calculate alpha for a single peer nomination item.

Combining multiple nomination items into a composite is one way to potentially improve the reliability of the data. If two items measure the same attribute, combining their results doubles *n*, as the additional item represents a new set of rater-item interactions. If the two items are positively correlated, the large increase in data will generally increase the measurement reliability. This study examined the potential gain in reliability from combining multiple limited nomination items by comparing reliability estimates from single-item measures with those from multiple-item composites.

Traditionally, researchers have calculated alphas for multiple-item composites by summing the number of nominations for each person for each question, and then calculating alpha across the questions (e.g., Becker & Luthar, 2007; Crick & Grotpeter, 1995). In this case, alpha indicates the extent to which the same participants are being nominated in aggregate for different aspects of a construct (e.g., for overt aggression, the extent to which children who are high on “hitting” would also be high on “kicking”). There is an alternate method to calculate alpha, based on the conceptualization of nominators as “items.” In this method, one takes the two 1/0 matrices and “pastes” them together, increasing the number of “item” columns, but keeping the number of nominees constant. One then calculates Cronbach’s alpha, treating each rater/rating combination as a unique item. In this context, alpha is essentially a measure of the extent to which nominators agree about which peers fit the nomination criteria as a whole.

These two methods of calculating alpha have the potential to yield different results. The “pasting” method preserves the nomination data in non-aggregated form. Using summed items for reliability estimation could yield differing results because of the change in the correlation structure of the data matrix. Such differences are discussed in the literature on split-half reliability, as the way that one splits an exam will affect the reliability estimate (Brown, 1910; Crocker & Algina, 1986; Spearman, 1910). For example, using the aggregated results of five items essentially estimates reliability using a “split-fifths” method. Previous peer nomination research has done little to compare reliability as estimated by these two methods. This study compared Cronbach's alpha values using the unique item/nominator combinations as items and using entire peer nomination items as items in the data setup.

Although most multiple-item peer nomination measurements *add* nominations together, some measurements require other mathematical operations to calculate a composite. For example, two items may be keyed in opposite directions. Opposite keying means that a nomination for one item indicates high relative standing on a trait, while a nomination for another item indicates low relative standing for the same trait. Some variations of Hartshorne and May's (1929) "Guess Who" instrument involved "positive" and "negative" nominations for behavioral criteria (e.g., Jarecky, 1959). In modern peer nomination research, this situation commonly occurs when measuring *social preference*. When people like or accept someone, the person's social preference standing is higher. When people do not like or reject someone, the person's social preference standing is lower. Acceptance and rejection are, therefore, oppositely keyed items measuring social preference. In addition to assessing the psychometrics of additive peer nomination items, this study examined social preference, calculated as the difference between acceptance and rejection.

Number of Nominators versus Participation Rate

Another important methodological issue for nomination reliability regards the participation rate versus the number of nominators (Marks et al., 2013). Before examining their nuanced implications for peer nominations, one must keep in mind that number of nominators and participation rate are highly positively correlated, especially when classroom size is homogenous across classes. It is our opinion that researchers should strive to reach the highest participation rate possible for numerous reasons, not the least of which are the problems that accompany missing data (Marks et al., 2013; Schafer & Graham, 2002).

Keeping this in mind, it is likely the raw number of nominators, not participation rate *per se*, is a primary driving factor in the reliability of peer nominations. The inter-rater correlation also plays a large role in the reliability of one's data. The associations among number of raters, the inter-rater correlation, and reliability can be seen in the equation for standardized alpha, which one would obtain when using nominator rating z-scores. Standardized alpha generally yields similar results to Cronbach's alpha (Equation 1). The equation for standardized alpha is

$$\alpha_z = \frac{n\bar{r}}{1+(n-1)\bar{r}}, \tag{2}$$

where $\bar{r}$ is the mean inter-rater correlation and *n* is the number of nominators (Lattin, Carroll, & Green, 2003). Distributing $\bar{r}$, Equation 2 becomes

$$\alpha_z = \frac{n\bar{r}}{1+n\bar{r}-\bar{r}}. \tag{3}$$

In order to understand this equation, it is useful to take limits to see what happens as *n* and $\bar{r}$ go to extreme values. Note first that if $\bar{r}$ is 0, alpha reduces to 0. The same occurs if *n* is 0, although this is the trivial case where one has no data.

Reliability will increase toward 1 as *n* or $\bar{r}$ increase. Assume that $\bar{r}$ is fixed. The limit as $n \to \infty$ is

$$\lim_{n\to\infty}\frac{n\bar{r}}{1+n\bar{r}-\bar{r}} = \lim_{n\to\infty}\frac{\bar{r}}{\frac{1}{n}+\bar{r}-\frac{\bar{r}}{n}} = \frac{\lim_{n\to\infty}\bar{r}}{\lim_{n\to\infty}\frac{1}{n}+\lim_{n\to\infty}\bar{r}-\lim_{n\to\infty}\frac{\bar{r}}{n}} = \frac{\bar{r}}{\bar{r}} = 1 .$$

As $\bar{r}$ approaches 1, 1 − $\bar{r}$ in the denominator cancels, leaving $n\bar{r}/n\bar{r}$, or 1.[1] A reliability estimate of 1, of course, indicates perfect reliability.[2]

The above exercise demonstrates that reliability increases as the number of nominators

[1] One can take these limits using Equation 1 and obtain the same results, but the mathematics are less obvious. We use standardized alpha for ease of illustration.

[2] Standardized alpha assumes that the mean inter-item correlation is positive. This equation fails when the mean inter-item correlation is negative.

and the average correlation between nominators increase. This study used the linear regression framework to model the associations among reliability estimates, nominator inter-correlations, number of nominators, and participation rate from real data to illustrate that the number of nominators has more impact on peer nomination reliability than participation rate.

The Current Study

This study had three goals. The first goal was to examine the potential increases in reliability when going from one limited peer nomination item to a composite of multiple items, using overt aggression, relational aggression, prosocial behavior, acceptance, and rejection as criteria. We hypothesized that single limited peer nomination items would lead to fairly unreliable scores, whereas multiple items would yield substantially more reliable scores. We also hypothesized that there would be a point of diminishing returns, whereby additional items would provide progressively smaller increases of reliability. This hypothesis is essentially stating that greatly increasing the number of ratings is similar to greatly increasing the number of items on an exam. Adding items to an exam generally increases exam score reliability, but there is a point where adding more items does not greatly increase reliability (Crocker, & Algina, 1986, Ch. 7).

The second goal was to investigate whether different peer nomination items differ in reliability. Using unlimited nominations, Marks et al. (2013) found that overt aggression scores were highly reliable, prosocial behavior somewhat less reliable, and acceptance (liking) the least reliable. The authors suggested that concrete observable behaviors, like overt aggression, should be nominated reliably, because participants can easily agree upon who engages in them. Individual affective reactions, like acceptance, should be less reliable because participants indicate their own personal feelings (cf. Babad, 2001). We expected our current results with limited nominations to mirror those of Marks et al.; that is, we expected overt aggression scores to be the most reliable variable, prosocial behavior to be less reliable, and acceptance to be substantially less reliable. We also assessed two variables not measured by Marks et al.: relational aggression and peer rejection (disliking). We expected relational aggression to be relatively high in reliability (although not as high as overt aggression, given that relational aggression is often covert; see LaFontana & Cillessen, 2002), and rejection to be about as reliable as acceptance.

The third goal was to outline two ways of estimating the reliability of multiple-item nomination composites: one using entire items as the basis of reliability estimation, the other using each nominator-item combination as the basis for reliability estimation. It is important to choose the best reliability estimation method for a study in order to have an accurate idea of the statistical integrity of the data. If the two calculation methods result in similar alphas, researchers can be assured that each method accurately estimates the reliability of multi-item measures. However, if the two methods yield different alphas, as we hypothesize, we can legitimately question whether we want to conceive the internal reliability of peer nomination measures as agreement on the level of the *nominators* or on the level of the sociometric *questions*.

In addition to these primary goals, this study sought to settle the argument on whether number of nominators or participation rate drives the reliability of peer nomination data. Based on the mathematical arguments outlined above, we hypothesized that the number of raters has a stronger impact on reliability than the participation rate.

Method

Data Source

Peer nomination data were collected from classrooms in 26 schools[3] in the American Midwest as part of a larger study concentrating on 470 focal participants. Although the larger

[3] An additional 2 schools (one classroom each) were excluded due to a loss of key data.

study was longitudinal, the current investigation focused on the data collected when focal participants were in the 6th grade. A total of 2,177 students were included as nominees on the peer nomination reports. The majority of students (92.1%) were in the 6th grade at the time of data collection; however, because of mixed-grade classrooms and other factors, students' grade levels ranged from 4th through 9th. The participants were 48.2% male and 51.4% female; all schools were co-educational. Participants were primarily White (64.3%) and African American (13.1%), with sizable minorities of Hmong[4] (3.8%), Asian American/Pacific Islander (3.2%), and Latino/Hispanic (2.3%) students. The socio-economic status of the schools' communities varied from lower to upper-middle class; in general, a large proportion of the students in this sample came from working-class backgrounds.

Consent documents were sent to parents of all students at participating schools. All students who returned signed consent forms also provided assent to participate; students were given candy and small, age-appropriate knickknacks for participating. Overall completion rates ranged from 25% to 96% across schools (*M* = 67.4%, *SD* = 17.1%); the raw number of nominators across schools ranged from 6 to 58 (*M* = 18.8, *SD* = 8.3). In total, 1,488 out of 2,282 students completed nominations.

The 26 schools included 77 classrooms in which data were collected. Depending on the educational structure of the school (i.e., whether the students in the target grade stayed primarily with one teacher or moved among different classes), students received a reference group list of either same-primary-classroom or same-grade peers of both genders. Because the definition of "classroom" varied across schools, class size (the number of participants on each nomination list)

[4] The Hmong people are a population originating from Southeast Asia. Significant enclaves of recent Hmong immigrants live in the American Midwest.

ranged from 19 to 109 (*M* = 28.9, *SD* = 14.6).[5] Lists were alphabetized with identification numbers next to each name; participants used these numbers when nominating peers for each item.

Research assistants verbally assured students of the confidentiality of their responses and read each question aloud as participants completed the nomination measure. Table 1 contains the peer nomination questions, along with abbreviations for the items that we used in tables and text.

Data Organization

Raw nominations were transcribed into a matrix format where every nominator (column) had a value of “1” for nominated peers (row) and a “0” for non-nominated peers.[6] This yielded a nominees by nominators matrix that was intentionally similar to the persons by items matrix psychometricians use to analyze achievement test data (Crocker & Algina, 1986). This data organization treats each peer nomination item as a test and each nominator as an item on that test. This data structure preserves the full pattern of nominations and was necessary to accurately estimate item reliability. Cronbach’s alpha was always calculated using Equation 1.

Combining Multiple Questions

We examined two ways to combine multiple peer nomination items into one measure for the purposes of estimating reliability. The “pasting” method took the nominee by nominator 1/0 matrices from multiple items and combined them by “pasting” them side-by-side. The number of

[5] It is worth noting that the varying definition of “classroom” means that the peer reference group differs for different participants. A participant who is allowed to nominate any peer in a grade is making a slightly different decision than a participant who is limited to nominating only peers in an immediate classroom (even though he or she might feel that peers in other same-grade classrooms better fit the nomination criteria). Differences between reference groups provide a relevant methodological issue (and/or limitation) in sociometric research, but the effects of these different reference groups are beyond the bounds of the current study (and not possible using the current data).

[6] Nominators always had a missing data value for the nomination of themselves. This planned missingness does not negatively affect the study (Schafer & Graham, 2002).

nominee rows stayed constant in this setup, but the number of columns, which represented unique combinations of nominators and items, increased substantially. The “items” method summed the nominations for each item and treated the sums for nominees as individual items. This approach resulted in a data matrix with much fewer columns (“items”), but the columns had much larger variance. We compared the reliability estimates that resulted from these two methods.

The difference between the two methods is important. In the “pasting” method, alpha indicates the extent to which the unique rater-item combinations agree, adjusted for the number of rater-item interactions. In the “items” method, alpha represents the degree to which the items in aggregate agree on individuals’ ratings, adjusted for the number of items. The two methods differ in that collapsing items into their aggregate sums changes the covariance structure of the items. In practical terms, the “pasting” method takes into account the relative agreement of individual raters, whereas the “items” method does not. If aggregating individual ratings into whole items appropriately raises the mean inter-item correlation, there will be little difference in the lower-bound reliability estimates between the two methods.

To test the first two hypotheses (regarding the reliabilities of single- vs. multiple-item measures), the reliabilities of multiple-item combinations were calculated using the “pasting” method of aggregation for overt aggression, prosocial behavior, and relational aggression (see Marks et al., 2013). In addition to investigating combinations of items with similar content keyed in the same direction, we also examined a composite of nominations keyed in opposite directions. The difference between nominations for acceptance and rejection yields a composite for social preference. This study calculated social preference using the “pasting” method by coding all nominations for rejection as -1 instead of 1. Descriptive statistics indicated if combining the variables could lead to a more reliable measure than the two variables alone.

Results

Reliability of Individual Items

Table 2 contains descriptive statistics for the reliabilities of the individual peer nomination items. There are several clear trends. First, the overt aggression scores were clearly the most reliable. The lowest mean overt aggression reliability estimate (.75 for Overt 2) was higher than all of the mean reliability estimates from the other item categories. Although we expected overt aggression to be more reliable than other items as stated in our second hypothesis, the fact that all five items reached "acceptable" levels of reliability in more than three-quarters of classrooms was surprising—we did not expect any single-item limited nomination measure to show such high reliability.[7] This high absolute level of reliability for overt aggression was contrary to part of our first hypothesis.

The prosocial behavior and relational aggression items showed similar reliabilities to one another ranging from .50 to .65, which is quite low. These reliability levels, compared to overt aggression, further confirmed our second hypothesis about the ordering of reliability levels and the type of variable measured. The standard deviations for reliability estimates were also consistent across these items. The exception was Rel 4, which had a lower mean reliability (.45) and a higher standard deviation (.30) than the other relational aggression and the prosocial items. There were also more classrooms with negative reliabilities for Rel 4. There was less agreement about which peers threaten to end friendships than about other relationally aggressive acts.

Table 2 also contains descriptive statistics for acceptance and rejection. These two items had extremely different reliabilities. Acceptance was by far the least reliable item in this study

[7] Three of the overt aggression items had a negative reliability estimate for a minimum value. These values were extreme outliers.

with a mean of only .16, while rejection was (contrary to the final expectation of our second hypothesis) the most reliable single item after overt aggression with a mean reliability of .68. Many classrooms had negative alphas for acceptance, as indicated by the first quartile being very close to 0. Thus, reliability for acceptance was very low, and Cronbach's alpha did not give a very good lower bound for the reliability when using limited peer nominations for this measure. Rejection, however, had no negative reliability values.[8] These two variables functioned quite differently than hypothesized.

Combining Multiple Items

The reliability estimates for most individual items were lower than desired. Past research has combined multiple items to obtain more reliable measures (Crick, 1997; Marks et al., 2013). Table 3 contains reliability statistics for combined items. Because the number of possible combinations is large, the table shows two combinations of *x* items for each variable set: the *x* most reliable items and the *x* least reliable items. For example, the two combinations of three overt aggression items were the three most reliable items (Overt 4 at .80, 1 at .80, and 3 at .78) and the three least reliable items (Overt 2 at .75, 5 at .76, and 3 at .78).

There are several interesting results. First, mean reliabilities were substantially higher for multiple items than for single items. This is best demonstrated by the second combination of items in each category (the two least reliable items). In each category, the combination of the two least reliable items was always more reliable than the most reliable single item (compare .87 to .80 for overt, .73 to .61 for Pros, and .68 to .65 for rel). This confirms part of our first hypothesis, which stated that multiple items will yield substantially more reliable scores than single items.

[8] Because the large difference in reliability between acceptance and rejection was surprising, we replicated this comparison using the dataset from Marks et al. (2013). Marks et al. did not originally analyze alphas for peer rejection. Mirroring the current results, the mean alpha for rejection ($\alpha = .82$) was significantly higher than the mean alpha for acceptance ($\alpha = .73$) across the 10 schools. Both values are higher than those in the current investigation because the previous dataset included higher *N*s and unlimited nominations.

Second, the lower part of the distributions of reliability estimates improved substantially when combining items. None of the minimum values for any combinations were less than 0. Many of the 1st quartile of reliability estimates were in the good (.80) or exemplary (.90) range.

Third, there was a point of diminishing returns for combining multiple items in the improvement in the mean reliability estimate. The gains in mean reliability were quite large when going from a single item to two items, and there was a modest gain when going from two to three items. The increases when going to four or five items were rather small. For example, combining the three most reliable Overt items yielded a mean alpha of .92; using the two additional items only increased the mean reliability to .95. This finding confirmed the final part of our first hypothesis.

Finally, combining acceptance and rejection into social preference (acceptance minus rejection) yielded a mean reliability estimate that was similar to the mean reliability of rejection alone (compare .68 for rejection with .65 for the combined variable). This is not surprising given the low reliability estimates for acceptance. Combining the items, however, substantially increased the minimum reliability estimate (−.99 and .05 for Accept and Reject respectively, .22 for Social Preference).

Calculating Alpha Using Items Versus Raters When Using Multiple Items

In order to compare the “pasting” and “items” methods of calculating Cronbach’s alpha, we calculated alphas according to both methods for the largest combination of items in each item grouping: 5 for overt and relational aggression, 3 for prosocial, and 2 for preference. The results of the two calculations were quite similar for highly reliable measures (e.g., overt aggression, r = .91; relational aggression, r = .84), but dissimilar for less reliable methods (e.g., prosocial behavior, r = .71; social preference, r = .27). The mean difference between the two calculation methods was nearly 0 for the aggression items. The “items” method gave slightly higher reliability estimates for the prosocial composite, but gave much lower estimates for social preference than using the rater-level data.

Participation Rate Versus Raw Number of Nominators

Finally, linear aggression analyses were used to examine the effects of the inter-rater correlation, number of nominators, and participation rate on Cronbach’s alpha reliability using the largest possible composite variable for each category. For each regression analysis, the classroom Cronbach’s alpha was the dependent measure. Each set of models began using only the mean inter-rater correlation to predict alpha, then added the number of nominators, then added participation rate as predictors. The change in R^2 can be compared across models. Statistical significance was set at $p < .01$ for the F test of significance of increase in model R^2.

Table 4 contains R^2 change analyses for predicting the dependent variable of Cronbach’s alpha for the largest composite variable from each nomination category using the independent variables of inter-rater correlation, number of nominators, and participation rate.[9] The first thing to notice in Table 4 is the dominance of the inter-rater correlation in predicting the reliability estimate by the size of the R^2 in the first (base) model for every category. Second, adding the number of nominators always resulted in a significant increase in R^2, increasing R^2 for Cronbach’s alpha between .20 and .35. Finally, participation rate never added for more than .05 to R^2. Participation rate resulted in a statistically significant increase in R^2 for prosocial behavior, but the increase in R^2 was only .04. Adding participation rate to a model predicting Cronbach’s alpha reliability did not substantially increase the prediction power above and beyond that of simply using inter-item correlation and number of nominators, which confirms our final

[9] To correct for curvilinear effects, we investigated several non-linear transformations to the variables. Although these transformations increased the R^2 values, the *relations* between the variables were nearly unchanged. Thus, we used the untransformed data for ease of interpretation.

hypothesis.[10]

Discussion

This study aimed to expand researchers' base of understanding regarding the methodological choices required when conducing peer nomination research, and the effects of those choices on the reliability of scores from these measures. Specifically, we investigated issues related to (a) optimal numbers of nomination items, (b) differing methods for calculating Cronbach's alpha for nomination data, and (c) whether researchers should be more concerned with *n* or participation rate when collecting data using peer nominations. These issues were investigated within the framework of a limited nomination procedure.

Our results were generally in line with our hypotheses, which stated that single-item measures would be unreliable, but combining multiple items would lead to substantially better reliability. First, several of the single-item measures were relatively unreliable. Although the single-item nomination measures for overt aggression had mean estimated reliabilities of .75 or greater, none of the other single-item measures had high estimated reliability. Even the overt aggression items had some classrooms with very low estimated reliabilities.

As expected from our first hypothesis, the reliability of the data improved substantially when combining multiple items into a single composite. Adding more items resulted in greater improvement in the reliability of the composite item, particularly for the least reliable classrooms. This improvement in the bottom areas of the reliability distribution occurred because of the number of very unreliable items a classroom can have. Although a classroom has a very unreliable result for a single item in a measurement set, it is quite unlikely (assuming some level

[10] The order in which the variables are added matters in the procedure used in Table 4. We also ran the analyses adding *part* before *n*. In all cases, *n* added much more to R^2 than *part*. We confirmed these results by running the largest models using the ANOVA framework. In all cases, the ANOVA *p*-values and η^2 effect sizes for *n* and *part* equaled the *p*-values and R^2 change values respectively from Table 4 to the number of decimals shown. This is not surprising, considering the mathematical relationship between the linear regression and ANOVA (Cohen, 1968).

of measurement validity) that a single classroom would have several very unreliable items. Combining multiple items, therefore, eliminates much of the chance of obtaining unreliable results at the classroom level.

There were diminishing returns to combining items: combining more than three items did not increase the mean reliability estimate much. This confirmed the final portion of our first hypothesis. The one caveat to this is with the minimum statistic. In all cases, the minimum classroom reliability estimate increased notably, even when going from four to five items. Additional items may have increased the minimum reliability even further. In order to guarantee that *all* measurements are highly reliable, a researcher should combine the highest number of limited peer nominations possible that still form a valid measure.

Second, this study replicated the general finding of Marks et al. (2013) that overt aggression is highly reliable, whereas peer acceptance is unreliable. This pattern of results supports the theory that nominations of concrete behaviors should be more reliable than nominations of individual affect. We formed our second hypothesis around this theory, and the data appear to support it. As expected, prosocial behavior and relational aggression, concrete behaviors not as visible as overt aggression, were between overt aggression and peer acceptance in terms of reliability. In fact, alphas for relational aggression and prosocial behavior were *very* similar when the number of items for each construct was the same.

The fact that limited nominations of peer rejection exceeded the “acceptable” level of reliability, on average, was contrary to our first hypothesis. We expected that rejection would show similar reliability to acceptance, given that the questions were nearly identical and tap similar constructs. There are a few explanations why rejection was more reliable than expected. It is possible that reliability is, in fact, affected by how concrete or consensus-based the construct is. Rejection may be tapping into a “group consensus” of who is disliked (just as popularity taps a group consensus of status; Cillessen & Marks, 2011), and is not based on individual-level affective reactions. It is possible that peers agree more on negative affective reactions (like rejection) than on positive affective reactions (like acceptance and friendship). These possibilities will have to be addressed in future investigations.

Third, this study examined the effects of estimating Cronbach’s alpha for a multiple-item composite using every unique person-rater combination as an item versus using entire items as items in the reliability calculation. While the method of estimation did not make a large difference when the composite had a high level of reliability, composites with lower reliability yielded very different reliability estimates with the two calculation methods. It appears that the covariance structure of lower-reliability nominations is more subject to unexpected fluctuations due to how the items are combined. We recommend that researchers use the finest (i.e., not summed or aggregated) data possible to compute reliability estimates. In this case, the finest data are each unique person-rater combination, which is seen in the “pasting” method. By using the individual ratings, the “pasting” method takes into account how much individual raters agreed when estimating reliability. The alternative method does not take this agreement into account at the individual level, because the reliability estimate aggregates out the individual ratings into entire questions. We feel that the “pasting” method is the best way to calculate the reliability of summed ratings, because relative rater agreement is an important aspect of the rating data. However, the final decision about whether to use the “pasting” or “items” method may come down to whether or not a researcher wants to explicitly account for rater agreement in the calculation of reliability.

Finally, this study examined the role of number of nominators versus participation rate in reliability estimates. The number of items (in this case, nomination item by rater interactions) and the mean inter-item correlation were the main predictors of Cronbach’s alpha, which confirmed our final hypothesis. This means that the acceptable participation rate for reliable scores may depend on class or grade size and the number of items a researcher combines into a composite. In small classrooms, it is important to obtain ratings from nearly everyone. Lower participation rates may be acceptable for larger classrooms or grades, particularly when combining several items into a composite.

These statements assume that the measures are valid. Missing data should not be taken lightly, and the above conclusions assume that missingness is completely at random (MCAR). Systematic missingness that a researcher does not properly address can lead to invalid measurement even if it appears to be reliable. The items that a researcher combines into a composite must also measure the same trait or set of traits. Combining heterogeneous items will also lead to invalid measurement. One must also keep in mind class or grade size when looking at participation. If a sample has relatively few nominators compared to nominees, there will likely be a large number of people without nominations. This problem is particularly acute for limited nominations. A higher participation rate is certainly a good way to increase the variation in summed ratings and decrease the floor effect due to many participants with zero nominations.

Obtaining the Most Reliable Peer Nomination Data Possible

Combining the results of this study with past research (e.g., Marks et al., 2013), we can recommend guidelines for obtaining the most reliable peer nomination data possible. First, it is more difficult to obtain reliable peer nomination data with limited nominations than with unlimited nominations. Having limited nominations restricts the variance of the variables. From a classical measurement perspective, items with proportions of endorsement between .4 and .6

maximize the reliability of a summed composite, as the higher variance of these items gives a greater chance at obtaining a high mean inter-item correlation and, thus, high item/total score correlations. It is almost always better to use unlimited nominations, as they will increase proportions of endorsement. The advantage is especially large when using single-item measures.

Second, combining multiple items measuring the same attribute will nearly always lead to more reliable measurement. When conceptualizing rater-item combinations as items in the Cronbach's alpha framework, going from one item to a composite of two items is the equivalent of doubling the length of a test. Longer tests generally yield more reliable measurement. This makes combining multiple items an extremely valuable tool in any researcher's toolbox.

The number of items to combine, however, depends on what statistical measures of reliability are important. If one cares most about the mean reliability of a variable, three items is a good number. One may want to combine five or more items in order to maximize every reliability estimate. The challenge is that all items involved truly measure the same construct. Future research should carefully examine how best to construct multiple peer nomination items that tap into the same trait. Of course, it may also be necessary to balance the need to maximize reliability and minimize the effects of fatigue—overly long measures may result in decreased reliability if participants stop paying attention or carefully considering their choices. This is likely to be particularly problematic when collecting data from younger participants.

Third, it is vital to obtain a sufficient number of unique nominator-item combinations. In other words, one should use enough items and raters so that, when combined, the number of columns in the rating matrix is high. With a single-item measure or a small classroom, a chance at reliable data may require a nominee participation rate of nearly 100%. The participation rate requirements may be lower when combining a large number of items, as long as the participation rate is not so small as to question the validity of the data in the first place.

Finally, the type of variable affects reliability. This and past studies (Babad, 2001; Marks et al., 2013) have indicated that nominations of easily observable behaviors, such as overt aggression, are more reliable. Nomination scores for affective measures are less reliable. Measures for which nominators may not systematically agree (e.g., friendship) will also be less reliable with measures like Cronbach's alpha. Peer nominations of these attributes will require multiple methods in combination to increase their reliability.

The preceding recommendations may be particularly important when conducting peer nomination research with young (i.e., preschool and kindergarten aged) participants, who we might expect to be less able to accurately or consistently provide "objective" information about peer behaviors and characteristics. For example, we know that young children tend to conflate status within the social hierarchy and peer acceptance, but that the relation between status and liking decreases as children get older and become better able to separate their peer reports from their friendships (Cillessen & Marks, 2011). Although we would certainly expect that multi-item and unlimited nomination measures would increase reliability for young children as well as older children, the ability to test the internal reliability of nominations with young children provides a unique opportunity to investigate the consistency and consensus of peer reports within preschool and kindergarten peer groups.

Limitations and Conclusion

As with any study, this study had some limitations. First, Cronbach's alpha was the single reliability estimate. This metric is justified, given that Marks et al. (2013) found that it works accurately in peer nomination analysis, and that the use of the primary alternative, Guttman's λ_2, showed nearly identical patterns of results. However, there is still some controversy in the literature regarding the appropriate ways to measure reliability, and future research could attempt to replicate these findings with different reliability metrics. Second, we examined peer nominations for only five types of traits. While we believe that the results from this study generalize to a wide variety of peer-nominated traits, it is important to test these results in additional contexts. Finally, this study did not address the validity of the multiple-item composites. Combining multiple items is certainly a good way to increase the reliability of nomination scores. It is of vital importance to research, however, that the individual items that compose the final product measure the same thing. Reliability is necessary but not sufficient for validity. High reliability indicates that one is measuring something well, but not necessarily measuring the correct thing.

A variety of sociometric research uses peer nominations. It is important that these measurements have high reliability so that researchers can properly describe the relationship between variables with as little interference from measurement error as possible. This research examined the case of limited nomination measures and suggested ways to obtain the most reliable peer nomination data possible. Using the tools suggested in this study, researchers can conduct improved peer nomination research in the future in order to gain a more complete picture of the relationship between a variety of sociological and psychological traits.

References

Babad, E. (2001). On the conception and measurement of popularity: More facts and some straight conclusions. *Social Psychology of Education, 5*, 3-29. doi: 10.1023/A:1012780232587

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,* 117-144. doi: 10.1111/j.1532-7795.2007.00514.x

Bronfenbrenner, U. (1944). A constant frame of reference for sociometric research: Part II: Experiment and inference. *Sociometry, 7,* 40-75. doi: 10.2307/2785536

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3,* 296-322. doi: 10.1111/j.2044-8295.1910.tb00207.x

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: Guildford Press.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin, 6,* 426-443. doi: 10.1037/h0026714

Coie, J. D., Dodge, K. A., & Coppotelli, H. (1982). Dimensions and types of status: a cross-age perspective. *Developmental Psychology, 18,* 557-570. doi: 10.1037/0012-1649.18.4.557

Crick, N., & Grotpeter, J. (1995). Relational aggression, gender, and social-psychological adjustment. *Child Development, 66,* 710-722. doi: 10.1111/j.1467-8624.1995.tb00900.x

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334. doi: 10.1007/BF02310555

Eng, E., & French, R. L. (1948). The determination of sociometric status. *Sociometry, 11*, 368-371. doi: 10.2307/2785197

Furr, R. M., & Bacharach, V. R. (2008). *Psychometrics: An introduction*. Thousand Oaks, CA: Sage.

Hartschorne, H., & May, M. A. (1929). *Studies in the nature of character II: Studies in service and self-control*. New York: Macmillan. doi:10.1037/11334-000

Jarecky, R. K. (1959). Identification of the socially gifted. *Exceptional Children, 25*, 415-419.

LaFontana, K. M., & Cillessen, A. H. N. (2002). Children's perceptions of popular and unpopular peers: A multimethod assessment. *Developmental Psychology, 38*, 635-647. doi: 10.1037/0012-1649.38.5.635

Lattin, J., Carroll, J. D., & Green, P. E. (2003). *Analyzing multivariate data*. Pacific Grove, CA: Brooks/Cole-Thompson Learning.

Liu, Y., & Salvendy, G. (2009). Effects of measurement errors on psychometric measurements in ergonomics studies: Implications for correlations, ANOVA, linear regression, factor analysis, and linear discriminant analysis. *Ergonomics, 52*, 499-511. doi: 10.1080/00140130802392999

Malcolm, K. T., Jensen-Campbell, L. A., Rex-Lear, M., & Waldrip, A. M. (2006). Divided we fall: Children's friendships and peer victimization. *Journal of Social and Personal Relationships, 23*, 721-740. doi: 10.1177/0265407506068260

Marks, P. E., Babcock, B., Cillessen, A. H. N., & Crick, N. R. (in press). The effects of participation rate on the internal reliability of peer nomination measures. *Social*

Development. doi: 10.1111/j.1467-9507.2012.00661.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Moreno, J. L. (1934). *Who shall survive? A new approach to the problem of human relations*. Washington, DC: Nervous and Mental Disease Publishing Company. doi:10.7326/0003-4819-8-1-104_2

Parkhurst, J. T., & Asher, S. R. (1992). Peer rejection in middle school: Subgroup differences in behavior, loneliness, and interpersonal concerns. *Developmental Psychology, 28*, 231-241. doi: 10.1037//0012-1649.28.2.231

Peery, J. C. (1979). Popular, amiable, isolated, rejected: A reconceptualization of sociometric status in preschool children. *Child Development, 50*, 1231-1234. doi: 10.2307/1129356

Rose, A. J., Swenson, L. P., & Waller, E. M. (2004). Overt and relational aggression and perceived popularity: Developmental differences in concurrent and prospective relations. *Developmental Psychology, 40,* 378-387. doi: 10.1037/0012-1649.40.3.378

Schafer, J. S., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147-177. doi: 10.1037/1082-989X.7.2.147

Spearman, C. (1907). Demonstration of formulæ for true measurement of correlation. *American Journal of Psychology, 18,* 161-169. doi: 10.2307/1412408

Spearman, C. C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271-295. doi: 10.1111/j.2044-8295.1910.tb00206.x

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. doi:

10.1002/cd.23220008805

Waldrip, A. M., Malcolm, K. T., & Jensen-Campbell, L. A. (2008). With a little help from your friends: The importance of high-quality friendships on early adolescent adjustment. *Social Development*, *17*, 832-852. doi: 10.1111/j.1467-9507.2008.00476.x

Wang, S. S., Houshyar, S., & Prinstein, M. J. (2006). Adolescent girls' and boys' weight-related health behaviors and cognitions: Associations with reputation- and preference-based peer status. *Health Psychology, 25*, 658-663. doi: 10.1037/0278-6133.25.5.658

Zeleny, L. D. (1940). Measurement of social status. *American Journal of Sociology, 45*, 576-582. doi: 10.1086/218376

Crick, N. R. (1997). Engagement in gender normative versus nonnormative forms of aggression: Links to social–psychological adjustment. *Developmental Psychology, 33*, 610-617. doi: 10.1037/0012-1649.33.4.610

Table 1
Peer Nomination Items and Abbreviations

Item	Abbreviation
Hits, kicks, punches others	Overt 1
Says mean things to insult others or put them down	Overt 2
Pushes and shoves others	Overt 3
Tells other kids that they will beat them up unless the other kids do what they say	Overt 4
Call others mean names	Overt 5
Does nice things for others	Pros 1
Helps others	Pros 2
Cheers up others	Pros 3
Tries to make other kids not like a certain person by spreading rumors about them	Rel 1
When mad, gets even by keeping the person from being in their group of friends	Rel 2
When mad at a person, ignores them or stops talking to them	Rel 3
Tells friends they will stop liking them unless friends do what they say	Rel 4
Tries to keep certain people from being in their group during activity or play time	Rel 5
Like	Accept
Don’t like	Reject

Table 2

Reliability Statistics for Individual Peer Nomination Items

Item	*M*	*Mdn*	*SD*	Minimum	1st Quartile	3rd Quartile	Maximum
Overt 1	0.80	0.85	0.16	−0.06	0.72	0.90	0.96
Overt 2	0.75	0.79	0.17	−0.13	0.69	0.86	0.95
Overt 3	0.78	0.85	0.19	0.03	0.73	0.90	0.95
Overt 4	0.80	0.84	0.18	−0.23	0.76	0.91	0.96
Overt 5	0.76	0.79	0.15	0.16	0.69	0.87	0.93
Pros 1	0.61	0.63	0.20	−0.07	0.49	0.78	0.88
Pros 2	0.50	0.55	0.21	−0.13	0.38	0.65	0.81
Pros 3	0.56	0.59	0.21	−0.36	0.49	0.72	0.82
Rel 1	0.65	0.72	0.20	−0.31	0.56	0.78	0.88
Rel 2	0.58	0.62	0.21	−0.07	0.51	0.72	0.88
Rel 3	0.52	0.58	0.22	−0.33	0.48	0.63	0.83
Rel 4	0.45	0.54	0.30	−0.52	0.37	0.64	0.86
Rel 5	0.55	0.59	0.20	−0.07	0.49	0.69	0.85
Accept	0.16	0.22	0.32	−0.99	0.07	0.34	0.63
Reject	0.68	0.72	0.17	0.05	0.58	0.80	0.95

Note. The unit of analysis in this table is the “classroom.” In other words, for a given item, half of classrooms showed reliability values above the listed median, half showed values below the median.

Table 3

Reliability Statistics for Combinations of Peer Nomination Items

Item	*M*	*Mdn*	*SD*	Minimum	1st Quartile	3rd Quartile	Maximum
Overt 4+1	0.89	0.92	0.09	0.53	0.87	0.95	0.98
Overt 2+5	0.87	0.89	0.09	0.38	0.83	0.93	0.97
Overt 4+1+3	0.92	0.95	0.07	0.61	0.91	0.96	0.98
Overt 2+5+3	0.91	0.93	0.07	0.56	0.89	0.95	0.98
Overt 4+1+3+5	0.94	0.95	0.05	0.73	0.92	0.97	0.98
Overt 2+5+3+1	0.93	0.95	0.05	0.71	0.91	0.96	0.98
Overt all	0.95	0.96	0.04	0.79	0.93	0.97	0.99
Pro 1+3	0.76	0.80	0.12	0.30	0.69	0.85	0.92
Pro 2+3	0.73	0.76	0.13	0.34	0.67	0.82	0.89
Pro All	0.82	0.84	0.09	0.51	0.78	0.88	0.93
Rel 1+2	0.77	0.81	0.12	0.34	0.73	0.85	0.94
Rel 4+3	0.68	0.71	0.14	0.26	0.58	0.78	0.91
Rel 1+2+5	0.83	0.86	0.08	0.55	0.78	0.88	0.95
Rel 4+3+5	0.77	0.81	0.11	0.42	0.71	0.85	0.93
Rel 1+2+5+3	0.85	0.87	0.07	0.57	0.81	0.90	0.96
Rel 4+3+5+2	0.83	0.86	0.09	0.49	0.79	0.89	0.96
Rel All	0.87	0.89	0.06	0.68	0.83	0.91	0.97
Social Preference	0.65	0.68	0.14	0.22	0.60	0.73	0.90

Note. The unit of analysis in this table is the “classroom.”

Table 4

Linear Regression R^2 Change Analysis for the Largest Composite Variable from Each Nomination Category

Model	Res. *df*	Res. *SS*	*df* change	*SS* change	*F*	*p*	Model R^2	R^2 Change
				Overt Agression				
$\alpha \sim r$	75	0.070	NA	NA	NA	NA	0.50	NA
$\alpha \sim r + n$	74	0.042	1	0.028	51.08	<.01	0.70	0.20
$\alpha \sim r + n + part$	73	0.040	1	0.002	3.05	0.08	0.71	0.01
				Prosocial Behavior				
$\alpha \sim r$	75	0.353	NA	NA	NA	NA	0.45	NA
$\alpha \sim r + n$	74	0.153	1	0.200	117.07	<.01	0.76	0.31
$\alpha \sim r + n + part$	73	0.125	1	0.028	16.29	<.01	0.81	0.04
				Relational Aggression				
$\alpha \sim r$	75	0.200	NA	NA	NA	NA	0.36	NA
$\alpha \sim r + n$	74	0.089	1	0.110	93.23	<.01	0.71	0.36
$\alpha \sim r + n + part$	73	0.086	1	0.003	2.51	0.12	0.72	0.01
				Social Preference				
$\alpha \sim r$	75	0.770	NA	NA	NA	NA	0.45	NA
$\alpha \sim r + n$	74	0.296	1	0.474	123.55	<.01	0.79	0.34
$\alpha \sim r + n + part$	73	0.280	1	0.016	4.24	0.04	0.80	0.01

Note: α is the Cronbach’s alpha by classroom (dependent variable), *r* is the mean inter-rater correlation by classroom, *n* is the number of nominators, *part* is the participation rate, Res. *df* is the residual degrees of freedom, and Res. *SS* is the residual sum of squares.