Supplemental Material

Task comprehension.
In our task, children's errors could stem from a cultural difference in their interpretation of "equal." Children may interpret "equal" as approximate equality rather than as exact equality. If this were true, then children should judge that only the take-half transformation breaks equality (as all other transformations either leave the set-size constant or change it by a cardinality of one, thus leaving their approximate size unchanged). However, only 4.76% (95% CI: 0.00-9.52%) of participants performed in accord with approximate size equality. This finding suggests that task performance reflects children's reasoning about exact equality, rather than any systematic misunderstanding of the question that was posed.

Influence of age and schooling on the exact equality task.
Although neither age nor years of education significantly correlated with performance on the exact equality task when the other was controlled (Taus < *0.08*, ps>*0.41*; both confidence intervals ranging above -0.12 and below 0.27; See main text), these variables were highly correlated (Tau=0.78), making them difficult to disentangle. Thus, we recomputed partial correlations controlling only for number word understanding. Exact equality task performance marginally correlated with age (Tau=*0.14*; *p=0.097*) and education (Tau=*0.14*; *p<0.102*).[1] This suggests that age, years of education, and/or some unmeasured variable that correlates with these two, influences children's performance in the exact equality task, independent of their knowledge of number words.

Predicted error rate of full counters.
If full counters' errors were due to temporary external distractions, then their error rate should be lower than the error rate of subset-knowers with similar age and schooling. To explore this idea we fit a linear model with the number of successful transformations as the dependent variable and age and years of education as the independent variables. The model was fit using data from subset-knowers. Next, we used the demographic information from full counters who did not perform at ceiling to predict their performance. Thus, the model's prediction reflect expected performance if the full-counters who erred on at least one transformations had a similar understanding of exact equality to that of subset-knowers. Supplemental table 1 shows the fit regression, and Supplemental table 2 shows the predicted scores for the twelve full counters who did not perform at ceiling. As Supplemental table 1 shows, children's performance was difficult to capture as a function of age and schooling. However, the model's predictions highly resembled the observed performance of the twelve full counters who did not perform at ceiling, suggesting that their errors were not substantially lower compared to subset-knowers, thus bolstering the claim that the errors of these full counters reflect a fragile understanding of exact equality rather than momentary distractions (see main text for main arguments).

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	2.54	0.15	16.91	<2e-16***

[1] These correlations become stronger when the aggregate score includes control trials.

Age (centered)	0.16	0.18	0.86	0.38
School (centered)	-0.11	0.31	-0.35	0.73

Supplemental table 1: Results from a linear regression with age and years in school as the dependent variable, and children's performance on the exact equality task as the independent variable.

Raw predictions	Rounded prediction	Observed performance
2.65	3	3
2.65	3	3
2.81	3	3
2.76	3	3
2.60	3	3
2.60	3	3
2.59	3	2
2.60	3	2
2.44	2	3
2.65	3	3
2.76	3	2
2.65	3	3

Supplemental table 2: Predictions using the linear model. Each row represents one of the full counters who did not perform at ceiling on the exact equality task. The left column shows the linear model's prediction if these full counters were performing similar to subset-knowers, the middle column shows the rounded prediction, and the rightmost column shows the observed value.

Relationship between each set transformation and knowledge of counting.
Overall, children's performance was highest on the stir transformation (*92.06*% success; 95% CI: 82.44-97.37). The take-half transformation (*87.30*% success; 95% CI: 76.50-94.35) was next highest, followed by the identity transformations (take and return and add and remove transformations combined, *79.37%* success; 95% CI: 67.73-88.53). The lowest success rates were on the addition and subtraction transformations (*73.02*% success each; 95% CI: 60.35-83.43) and the substitution transformation (*65.08*% success; 95% CI: 52.03-76.66).

In the main text, our analysis focused on children's understanding of exact equality as given by their aggregate performance in the exact equality task. However, there may be more intricate patterns in each individual transformation. We explored this possibility in a mixed effects logistic multiple regression designed to explore how counting knowledge relates to each individual transformation. In this regression we predicted individual trial performance on each individual set transformation from the transformation type, the child's number-knower level, and age, with by-child random intercepts. For each of the independent variables we chose the coding scheme that would best allow us to interpret

the coefficients. First, set transformation type was deviation (a.k.a sum) coded, such that the mean performance of each transformation was compared to the overall mean performance (across transformations). Number-knower level stages were represented numerically (range from -3 to 1, with 1=full-counter), with interactions by set transformation. Age was standardized (i.e., z-scored) and entered as a control predictor. Overall, the regression simultaneously fits coefficients for the (a) overall improvement in set performance as a function of children's knower levels, (b) differences in performance among set tasks, (c) age, and (d) interactions between knower-level and set transformations (such that different knower-level stages may influence performance differently for each transformation). The regression revealed a coefficient of 0.42788 for each standard deviation in age, which is equivalent to a coefficient of 0.25 per year when age is only centered. The full regression table is shown in Supplemental table 3.

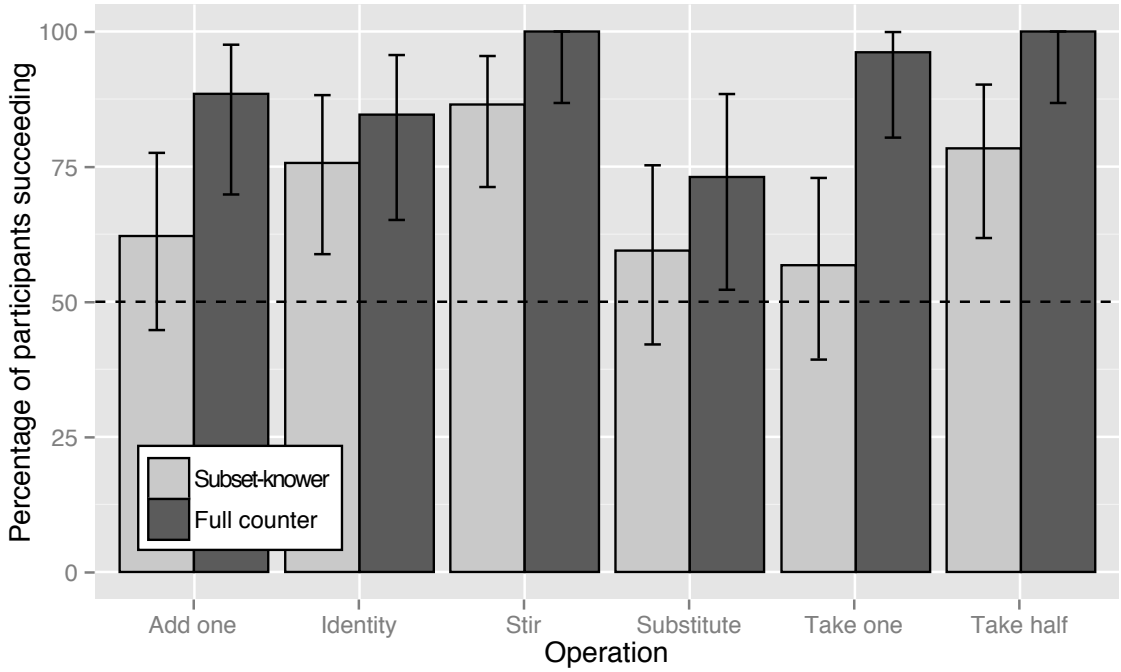
The regression was consistent with our main analysis. Overall, children performed better as they grew older (Beta=0.25 per year, z=2.1; $p<0.05$) and as their knowledge of counting improved (i.e., as they progressed through each knower-level stage; Beta=0.4 per level, t=3.20; $p<0.05$). Additionally, the regression suggested that children in lower knower-level stages made mistakes across all transformations, but as their knower-level increased, their errors significantly concentrated in the substitution transformation (Beta=-0.53; t=-3.04; $p<0.01$ on the interaction between knower-level and the substitute transformation; set-transformations were sum coded). Thus, children did not improve on all transformations uniformly, but rather had more trouble understanding how substitutions affect a set's size.

Next, we looked at children's performance in each transformation as a function of their ability to count. As Supplemental Figure 1 shows below, children performed above chance in the stir, identity, and take-half transformations, independent of their ability to count. In contrast, only children who could count performed above chance in the add one, take one, and substitution transformations. These results are consistent with the hypothesis that children are learning to understand exact equality, and with previous data from 3-year-old children in industrialized societies (see Izard, Streri, & Spelke, 2014). The stir and identity transformations maintain the set's exact size, but also the exact elements in the sets, and the take half transformation changes the cookie pile's shape enough to make an easy judgment using geometrical cues. In contrast, the add one, take one, and substitution transformations change the elements in the set, and yet produce only subtle changes in geometrical cues, as each pile of cookies was rearranged to highly overlap, making it difficult to see if there were any other elements hidden under the pile.

	Estimate	**Std. Error**	**Z value**	**Pr(>|z|)**	***
(Intercept)	1.99883	0.24227	8.25	<2.00E-16	***
Knower level	0.40174	0.12566	3.197	0.00139	**
Stir	1.00678	0.62889	1.601	0.1094	
Add	-0.68549	0.38341	-1.788	0.07379	.
Take	-0.13713	0.45965	-0.298	0.76544	
Identity	-0.52617	0.38591	-1.363	0.17274	

Substitute	-1.43175	0.33823	-4.233	2.31E-05	***
Age	0.42788	0.20414	2.096	0.03608	*
Knower level : Stir	-0.03522	0.2762	-0.128	0.89852	
Knower level : Add	-0.10117	0.18876	-0.536	0.59199	
Knower level : Take	0.34342	0.22496	1.527	0.12686	
Knower level : Identity	-0.31819	0.19313	-1.648	0.09944	.
Knower level : Substitute	-0.5311	0.17496	-3.036	0.0024	**

Supplemental table 3: Results of a mixed effect logistic multiple regression predicting children's performance on the task (success or failure). Set transformation type was deviation (a.k.a. sum) coded, and number-knower levels were represented numerically (0-, 1-, 2-, and 3-knowers, as -3,-2,-1, and 0, respectively, and full counters as 1), and age was standardized (z-scored). The Knower-level : Transformation interaction was included through an analysis of deviance ($p<0.005$).



Supplemental figure 1: Children's performance on each of the six set-transformations. The light gray bars indicate participants who have not yet learned to count and the dark gray bars indicate participants who can count. The solid vertical lines show 95% confidence intervals and the horizontal dotted line is the expected chance behavior.