Published in final edited form as:

J Exp Child Psychol. 2009 July ; 103(3): 325–337. doi:10.1016/j.jecp.2009.02.007.

Levels of Number Knowledge in Early Childhood

Barbara W. Sarnecka and Michael D. Lee

Department of Cognitive Sciences, University of California, Irvine.

Abstract

A task commonly used to assess young children's number knowledge is the Give-N task, in which an experimenter asks the child to generate sets of a given number. Typically, children find small sets easier to produce than large ones. The present study analyzes children's incorrect responses (i.e., errors) on the Give-N task. Logically, there are two types of incorrect response possible: (a) performance errors, and (b) guesses. Performance errors occur when the child knows the meaning of the number word requested, but makes a mistake in producing the set. Guesses occur when the child does not know what the number word means. New analyses of three previously published datasets (280 children total) found that most errors were guesses, with some evidence of performance error appearing for high number words only. Results support the “knower-levels” theory of number development.

Children learn to count (i.e., to recite the number-word list) very early. But how can we (adults, researchers) know whether children understand the cardinal meanings of the number words? Some would argue that a child who recites the number-word sequence, and certainly a child who counts objects correctly, has demonstrated that she understands the cardinal meanings of the number words. Others believe that children fail to understand the cardinal meanings until much later. This disagreement has led to the invention of a variety of different tasks probing children's understanding of number-word meanings.

One commonly used task is the Give-N task (e.g., Frye, Braisby, Lowe, Maroudas & Nichols, 1989; Fuson, 1988; Schaeffer, Eggleston & Scott, 1974; Wynn, 1990, 1992.) In this task, the experimenter asks the child to generate a set of a particular size. For example, the child might be asked to “give three bananas” to a stuffed monkey or to “put five fish” in a bowl. Children find this task more difficult than simple counting, and the larger the set they are asked to produce, the more difficult it is. (Condry & Spelke, 2008; Le Corre, Van de Walle, Brannon & Carey, 2006; Le Corre & Carey, 2007; Sarnecka & Carey, in press; Sarnecka & Gelman, 2004; Sarnecka, Kamenskaya, Ogura, Yamana & Yudovina, 2007; Schaeffer et al., 1974; Wynn, 1990, 1992).

It probably surprises no one that children make more errors with larger set sizes than small ones; but we do not actually know why this is so. Logically, there two broad types of error possible. The first is systematic performance error in counting and/or estimation. This type of error occurs when the child understands the meaning of the number word, but makes mistakes producing the set. The second type of error is guessing. This occurs when the child does not understand the cardinal meaning of the number word asked for, and thus does not know how many objects to give. Let us consider each type of error separately.

Performance Error

A child who is asked for “five” items, and understands that “five” means 5, can still make mistakes producing the set (Cordes & Gelman, 2005; Gelman, Meck & Merkin, 1986; Greeno, Riley & Gelman, 1984). If children attempt to count out the required number of items, their errors will show what has been called binomial variability (Cordes, Gelman, Gallistel & Whalen, 2001). If children try to estimate the correct number, their errors will show scalar variability (Dehaene, 1997; Gallistel, 1990; Platt & Johnson, 1971; Whalen, Gallistel, & Gelman, 1999).

The response curves predicted for each type of error are illustrated in Figure 1. Notice that counting and estimation error have a lot in common. First, the distribution of incorrect answers for any target number word will have the same mean as the target number itself. For example, if we look at all trials requesting “five” items, and we take out all of the correct responses, the mean of incorrect responses should still be 5.

This is because counting and estimation error are both symmetrical—under-counting happens with the same frequency as over-counting; underestimation is as frequent as overestimation. For any given target, counting and estimation errors fall equally above and below it. If the child is asked for “five” objects, then responses of 4 and 6 are equally frequent; responses of 3 and 7 are equally frequent; responses of 2 and 8 are equally frequent; and responses of 1 and 9 are equally frequent.

Another similarity between counting and estimation error is that errors close to the target happen more often than errors far away from it. Thus, if the child is asked for “five,” then responses of 4 and 6 will be the most frequent errors, followed by 3 and 7, followed by 2 and 8, etc. In other words, counting and estimation errors both cluster symmetrically around the target.

For both counting and estimation error, the distribution is more spread out around the higher numbers. This is how performance error accounts for the fact that children perform better with small numbers than large ones. In the case of counting, this is because longer counting sequences offer more opportunities for mistakes. In the case of estimation, it is because estimation error is proportional to the size of the target. (As the numbers get higher, the distribution for estimation error is *more* spread out than the distribution for counting error. This is the difference between scalar and binomial variability.) For our purposes here, however, the similarities between counting and estimation error are more important than the differences: Both counting and estimation produce symmetrical distributions of errors which are centered on the target.

(Educated) Guesses

A totally different type of incorrect response is the guess. Many scholars have argued that even after children have learned to count, both intransitively (reciting the number-word list) and transitively (counting objects), they still take a long time to learn the cardinal meanings of the number words (e.g., Carey, 2001, 2004; Fuson, 1988; Le Corre & Carey, 2007; Le Corre et al., 2006; Sarnecka & Carey, in press; Sarnecka & Gelman, 2004; Sarnecka et al., 2007; Wynn, 1990, 1992).

According to this view, most of children's errors on the Give-N task reflect genuine ignorance of the number words' cardinal meanings. Children succeed first at small numbers because they learn the number-word meanings one at a time and in order-- first “one,” then “two,” then “three”-- over a period of many months. This view is called the “knower-levels” view, because children who have learned the meaning only of “one” are called “one”-

knowers; children who have learned “one” and “two” are called “two”-knowers, and so on (Wynn, 1990, 1992; see also Le Corre, et al., 2007).

According to the knower-levels view, after the child has learned the words “one,” “two,” “three” (and perhaps “four,” depending on the individual), she figures out how counting works. Now she can assign meanings to the higher number words via counting— this is how the cardinal meanings of the number words “five” and higher are learned. Once the meanings of these number words are understood, then the child's errors on the Give-N task will no longer be guesses, but counting and/or estimation errors. Before the words' meanings are understood, however, errors reflect guessing—in other words, responding at chance.

But what is ‘chance’ responding in this task? One easy assumption would be that the distribution is flat (i.e., that all responses from 1-15 are equally likely). This would make sense if children's responses were spun on a roulette wheel. But given the practical demands of the Give-N task, some default responses seem more likely than others. Responses that involve only a single gesture-- such as giving one item, giving a handful of items, or pushing the whole pile of items across the table-- all seem more likely than responses that require multiple gestures (i.e., giving more than one handful, but less than the whole pile.)

Aside from the difficulty of specifying the chance distribution for this task, there is a further complication. The knower-levels view predicts one important constraint on children's guesses: When a child is asked for a number word whose meaning she doesn't know (e.g., when a “three”-knower is asked to give five objects), she should avoid giving any number that she *does* know (e.g., a “three”-knower will not give 1, 2 or 3 objects). This is because children treat the number words as mutually exclusive (Wynn, 1992).

The knower-levels view predicts that the child will generate sets correctly for the numbers that she knows (i.e., the numbers up to and including her knower-level). For the numbers above her knower-level, she will respond at chance (which is probably not a flat distribution for this task), but she will avoid giving any number that she does know. For example, a “four”-knower who is asked for “five” objects might give 6, 7, 8, or 9 objects, but she will not give 4, 3, 2, or 1, because she knows what those cardinalities are called. This prediction is in direct conflict with that of performance error, where the child is assumed to know what all the words mean, and trials asking for 5 objects should elicit errors of 4 and 6 with equal frequency; 3 and 7 with equal frequency, etc. Thus, the knower-levels theory doesn't predict wild guesses, but educated guesses.

Are Knower-Levels a Real Phenomenon?

Many previous studies have used the knower-levels framework (e.g., Condry & Spelke, 2008; Le Corre et al, 2006; Le Corre & Carey, 2007; Sarnecka & Gelman, 2004; Sarnecka et al., 2007; Wynn, 1990, 1992). Several of these (Le Corre et al, 2006; Le Corre & Carey, 2007; Wynn, 1992) reported high levels of agreement between a child's knower-level based on the Give-N task and her knower-level as assessed via some other task. Wynn (1992) compared children's Give-N performance to their performance on the Point-to-X task, where the child is shown two cards and asked to point to the card with (for example) “three” fish. Wynn reported nearly 100% agreement between Give-N knower-levels and Point-to-X knower-levels. More recently, Mathieu Le Corre and colleagues (Le Corre et al, 2006) compared children's Give-N performance to their performance on the What's-On-This-Card task (Gelman, 1993), where children are shown a card picturing some set of objects. The study reported high levels of agreement between the two measures. For both Wynn and Le Corre, inter-task consistency was evidence that the knower-levels framework captures a real developmental phenomenon.

These findings are convincing to readers inclined to accept the knower-levels framework, but less convincing to those who may be skeptical of the existence of knower-levels in the first place. Studies that begin by sorting children into knower-levels are *ipso facto* unable to prove that knower-levels are (or are not) a real phenomenon, for several reasons. First, as we have already shown, the prediction that children should perform better when asked for lower numbers is consistent with both performance error and guessing.

Second, the data in most knower-level studies are coded in a way that makes it impossible *not* to find knower-levels. For example, Sarnecka and colleagues (Sarnecka & Gelman, 2004; Sarnecka et al., 2007) gave children three trials for each number word. In order to be considered successful at that number word, the child had to get 2 of those 3 trials correct, and also give the target number no more than once as a response for any other number word. (For example, a child who always gave 2 when asked for “two,” but also gave 2 when asked for “three,” was considered *not* to know “two”.) The child’s highest successful number became her knower-level.

When (1) coding practices make it inevitable that knower-levels will appear in the data, (2) different tasks can be solved by the same strategies (i.e., counting or estimation), then it is perhaps not surprising that many children score at the same knower-level across different tasks. A constant level of performance (counting or estimation) error for each child would produce success at the same low numbers and failure at the same high numbers across tasks.

The Present Study

What is needed, then, is an analysis of Give-N errors that can distinguish between performance errors and guessing. The present study provides such an analysis. We organize this analysis around two questions.

The first question is whether children’s errors show the characteristic, symmetrical pattern of performance error. If either counting or estimation error is the principle source of error in these data, then incorrect responses should cluster around the target number, falling symmetrically above and below it, so that the mean incorrect response for each target number is equal to the target number itself. Alternatively, if most of the errors are guesses, they need not be related to the target at all.

The second question concerns the knower-level constraint on children’s guessing. (This is what would make the guesses ‘educated.’) According to the knower-levels view, a child should not give a number she knows, when she is asked for some number word she doesn’t know. To understand how this differs from counting or estimation error, imagine two scenarios in which you (the reader) might be a subject in an adult Give-N study.

In the first scenario (which illustrates the predictions of counting/estimation error) you are given four Give-N trials: On three of the trials, you are told to give 99 bananas to the monkey, and you do correctly. On the fourth trial, you are told to give 100 bananas, but you make two mistakes while counting out the set. So how many bananas do you actually give? There is an equal probability that you give 98 bananas (if you over-counted both times) or 102 bananas (if you under-counted both times). Or you might give 100 bananas if you made one mistake of each type (but then your answer would be correct and no one would be the wiser). The point is that your chances of giving 98 are the same as your chances of giving 102. It doesn’t matter that you already gave 99 items successfully. There’s no reason that your error (i.e., 98) shouldn’t fall below some previous correct answer (i.e., 99).

Now imagine a child doing the same task, but instead of 99 items, she is asked to give 4 (and does so correctly). Then, instead of 100 items, she is asked to give 5 (and makes two

mistakes). She might end up giving 3 items (if she over-counted both times), or 7 items (if she under-counted both times). But the likelihood of these two responses is exactly equal, because 3 and 7 are both equally different from the target 5. The fact that she previously gave 4 items without error has nothing to do with it.

To understand how the knower-levels explanation is different, imagine a second Give-N scenario. In this version, you are again a subject in a Give-N study, but the number-word requests will be in Japanese. Imagine, moreover, that you know a little Japanese, but not much. The only number words you know are *hitotsu* (“one item”), *futatsu* (“two items”), *mitsu* (“three items”) and *yotsu* (“four items”). Whenever the experimenter asks you for one of these number words, you give the correct number of items. But then the experimenter asks you for *itsutsu*. You don't know how many that is, so you have to guess. But you can be fairly confident that *itsutsu* does not mean one, two, three or four items, because you already know the words for those numbers. So you will make an educated guess, and give 5 items or more. Now, it just so happens that *itsutsu* means “five items.” So maybe you get lucky and gave 5. But if not, your errors will be asymmetrical, because you might give more than 5, but you will not give less. So, you will give 6 but not 4; 7 but not 3; etc. This asymmetry is in direct conflict with the symmetrical pattern of performance error.

The goal of the following analyses, then, is to characterize the nature of children's incorrect responses in these data. Are they same sort of predictable, symmetrical errors of counting and estimation that we find in adult data? Or are they educated guesses— constrained by the numbers that children do know, but otherwise reflecting some (unspecified) chance distribution?

Method

Dataset 1

These data were collected from 82 monolingual speakers of English, ages 2-4 years (mean 3 years, 7 months; range 2-11 to 4-8), tested at preschools in Irvine, California, or at a university cognitive development lab in Cambridge, Massachusetts. As part of their participation in other studies, each child was tested on two standard counting tasks: intransitive counting (i.e., reciting the number-word list up to “ten”), and transitive counting of arrays of 5 and 10 objects. Every child produced the English number-word list up to “ten” on the intransitive counting task (mean score 10.0), and nearly every child counted 10 objects correctly on the transitive counting task (mean score 9.6). Our analysis focuses on these children's performance (specifically, their errors) on the Give-N task.

For the Give-N task, the child was presented with a bowl of 15 identical, small toys (e.g., rubber bananas approx. 3 cm long) and was asked to place some number of the toys on a plate for a stuffed animal. (E.g., “Can you give *three* bananas to the monkey?”) After the child finished placing objects on the plate and slid the plate across the table to the animal, the experimenter asked a single follow-up question, of the form, *Is that N?* (E.g., “Is that *three*?”) If the child said no, then the experimenter returned the plate to the child and restated the original prompt (e.g., “Well, the monkey really wants *three* bananas. Can you give him *three*?”) This continued until the child affirmed that the animal had received the right number of objects. Each child completed 4-21 trials of this task (mean 12.4 trials). Numbers requested included “one,” “two,” “three,” “four,” “five,” “six,” “eight” and “ten.” For details of the procedure, see Sarnecka & Carey (in press).

One quirk of the data collection was that the number “six” in Dataset 1 was only requested of children who had correctly given 5 when asked for “five” on a previous trial. Children who had not given 5 correctly were never asked “six” in this dataset. This produced a

somewhat different distribution of errors for “six” than for other target numbers (see *Results*, below). Trials for “six” in Dataset 1 were the only subset of data affected in this way.

Dataset 2

These data were collected from 162 children, ages 2-3 years (mean 3 years, 2 months; range 2-9 to 3-7), living in Ann Arbor, Michigan (70 children); Kobe, Japan (48 children); and St. Petersburg, Russia (44 children). All children were tested in their native language by a female native speaker.

Each child completed 9-15 trials of the Give-N task (mean 14.7 trials). The numbers requested were “one,” “two,” “three,” “five” and “six.” These children were not given a count-to-ten task. For details of the procedure, see Sarnecka et al. (2007).

Dataset 3

These data were collected from 36 children, ages 2-4 years (mean 3 years, 3 months; range 2-7 to 4-1), tested in communities around Boston, Massachusetts. Of these, 21 were native English speakers and were tested in English; the other 15 were native Russian speakers, tested in Russian.

Each child completed 26-36 trials of the Give-N task (mean 35.6 trials). The numbers requested were “one,” “two,” “three,” “four,” “five,” and “ten.” These children were not given a count-to-ten task. For details of the procedure, see Sarnecka, et al. (2007).

Results

It seems desirable to pool the data from all three datasets in order to make the analysis maximally powerful. On the other hand, performance error is only possible if children know the number words, and this could raise questions about Datasets 2 and 3, where children were not asked to count to ten. (Many people have argued, of course, that being able to count is no guarantee of knowing what the number words mean. But children must at least be able to produce the list, before it makes sense to ask whether they know what the words mean.) Thus, each of the following analyses was conducted twice: Once for all three datasets together, and once for Dataset 1 by itself.

First question: Do errors cluster symmetrically around the target, such that the mean incorrect response for each target is equal to the target number itself?

The mean incorrect response for each number word is given in Table 1; the data are plotted in Figures 2a and 2b. For the pooled data, the mean incorrect response was significantly different from the target number in every case but one. The exception, “eight” was entirely comprised of data from Dataset 1 and is discussed below. Thus, for the pooled data, there is no support for the hypothesis that counting or estimation error was the principle source of variation. Most of the errors really do appear to be guesses.

For Dataset 1 alone, where children were older and more skilled at counting, the picture is slightly more complicated. For the lower number words (“one” through “five”) the mean incorrect responses were all significantly different from the target number ($ps < .001$), just as in the pooled data. But for the high target words “six,” “eight” and “ten,” the mean incorrect responses were not significantly different from the target. This could be explained in one of two ways. Either (a) most errors for low-number targets were guesses, but most errors for high-number targets were counting/estimation errors, or (b) most of the errors for all targets were guesses, and the mean of these guesses happened to fall around 8.

In the case of target number “six,” the first explanation seems plausible, because of the quirk in the data collection for “six” in Dataset 1 (see *Method/Dataset 1*, above). In Dataset 1, the only children asked for “six” items were those who had given 5 objects on a previous trial asking for “five.” As Figure 2b shows, many of these children gave 5 objects for “six” as well (indicating perhaps that they didn't know what “five” or “six” meant, but gave 5 objects as an educated guess for words they didn't know). But the fact that most of the errors for “six” involved responses of 4-8 objects suggests that many of these children did know what “six” meant, and simply made mistakes in counting or estimation.

In the case of “eight” and “ten,” the second explanation seems more plausible. That is, most of the errors seem to be guesses, and the mean happens to fall around 8. In this dataset, the mean incorrect responses for target words “four,” “five,” “eight” and “ten” were not significantly different from each other (one-way ANOVA $F=1.11$, $p = .30$, N.S.). If children give approximately 8 items regardless of whether they are asked for “four,” “five,” “eight” or “ten,” it would seem wrong to conclude that they know what “eight” means.

A similar phenomenon occur in the pooled data, for the target numbers “four” and “five.” Because the mean incorrect response for most target numbers in the pooled data was in the range of 4-5, the mean appeared to be close to the target only for those words. But in fact, children's answers changed little from target to target, and were not even correlated with the target word ($r=.16$ for the pooled data; $r = .31$ for Dataset 1 alone).

The data plots reinforce this impression. For both the pooled data and Dataset 1 alone, similar error distributions appear across very different target numbers. Responses of 2, 3 and 4 are very common, as is 15 (i.e., giving all the objects). It is noticeably rare for children to give large numbers of items that are less than the whole pile (i.e., responses in the range 9-14 are relatively rare.) Even excluding the many responses of 15 (which are clearly guesses) errors appear to be clustered symmetrically about the target only when the target happens to fall in smack in the middle of the recurring distribution.

To put this another way, consider the response curve for “five” in Dataset 1. If we ignore the many responses of 15 (clearly guesses), we are left with a fairly symmetrical distribution around the target number “five.” But the errors for target number “four” are also clustered around “five.” So are the errors for “eight” and the errors for “ten.” So when we look “five” in the context of all the other data, the most likely explanation seems to be that the distribution of children's guesses in this dataset happened to center around “five”

In spite of this, there are hints of at least a few counting or estimation errors. The distribution of errors for target number “ten” in the pooled data includes small bumps at 9 and 11 that do not appear in the distributions for other target words. This indicates that at least a few of the children knew what “ten” meant, and were trying to estimate or count out 10 objects. Similarly, as mentioned above, the distribution for “six” in Dataset 1 looks different from all the other distributions. This group of children who had given 5 correctly were different from the larger subject pool. None of them guessed that “six” meant 15, and none of them guessed that it meant 1, 2 or 3. Of all the distributions pictured, this is the one that comes closest to satisfying the predictions of performance error.

Thus, the bulk of the errors in our data appear to be guesses, and their distribution is similar from one target number word to the next. This leads us to a problem. If most guesses are in the range of 2-5 objects, then how can we tell the difference between guesses and performance errors for the target numbers “two” through “five”? Any performance errors would simply disappear in the mountain of guesses that fall in the same range. The answer lies in our second analysis.

Second question: If a child knows the meaning of a given number word (e.g., “four”), and is asked to produce a higher number (e.g., “five”) does she ever produce in error a set that is lower than the number she knows (i.e., 1, 2, or 3)?

This analysis takes advantage of the knower-level theory's prediction that children's educated guesses should be lower-bounded (they should not give numbers they can identify), whereas performance error is symmetrical.

First, we identified the group of children who performed perfectly at a given number. For example, for “four,” we identified the subset of children who always gave 4 objects when asked for “four,” and never giving 4 objects otherwise. (Note that this says nothing *a priori* about how the children performed at lower or higher number words.) Then we looked at the mistakes made by that group of children when they were asked for higher number words (e.g., for the group perfect at four, we looked at their errors on number words “five” and above.)

Figure 2 illustrates an example. The pair of numbers under scrutiny are four and five, with four being the number children were perfect on (indicated by a “P” on the graph), and five being the number they were asked for (indicated by an “A” on the graph). The X-axis lists all the possible set sizes children could produce; the Y-axis shows the number of times each error was produced. Naturally there are no errors of 4, because this graph shows data from the subset of children who never gave 4 objects in error. Responses of 5 do not appear because they wouldn't be errors—five was the number asked for. White columns on the graph show errors that were not counted in our analysis. These include errors of 6, which were not counted because errors equally far from the target in the other direction (i.e., errors of 4) had been selected out. Responses of 10 and above were not counted for the same reason—it was not possible to make equidistant errors in the other direction (i.e., these would have been responses of 0 or less). That left three pairs of possible errors to be compared—3 vs. 7; 2 vs. 8; and 1 vs. 9. For two of these pairs (3 vs. 7; 1 vs. 9) this example contained no errors and thus provided no data. For the pair 2 and 8, however, there were two errors. Both of them were responses of 8. So the final count in this example is 2 errors above the target; 0 errors below it. This analysis was repeated for all possible pairs of numbers in all three datasets.

There is one more detail to note: Although we excluded responses that were too high to have equidistant errors below the target (e.g., we excluded errors of 10 and above in the previous example), we actually *included* responses that were too *low* to have equidistant errors above the target (because these might provide important evidence against the knower-levels explanation.) For example, if a child was asked for “ten” items and gave only 2, we included that in our count. We did so even though an equivalent error in the other direction (i.e., a response of 18) was impossible. In other words, we structured the analysis in a way that was actually biased against the knower-levels explanation.

The final counts for each dataset were as follows: Dataset 1 contained 28 errors above the target (one or more for 19 of the 82 children) and 0 errors below it. In other words, Dataset 1 contained no errors that fell below a number the child demonstrably knew.

Dataset 2 contained 31 errors above the target (one or more for 17 of the 162 children) and 1 error below it. This error was made by a child who performed perfectly on two, but gave 1 object on a single trial asking for “three.”

Dataset 3 contained 11 errors above the target (one or more for 6 of the 36 children) and 12 errors below it, all produced by the same child. This child performed perfectly when asked for “one,” “two,” “three,” and “four,” but gave 2 objects in all 12 trials asking for “ten.”

Because this behavior does not fit the predictions either of performance error or the knower-levels framework, we conclude that the child must have misunderstood the instruction, perhaps mistaking the word “ten” for the word “two.” Aside from this child, there were no other below-target errors in Dataset 3.

The grand total, then, was 13 below-target errors, and 70 above-target errors, a clearly asymmetrical result, $p < .001$. Thus, we find very little evidence of counting or estimation errors buried in the mountain of guesses. It is still possible that the data contain a few such errors. Our analysis included only children who performed perfectly at a given number, so a child who made performance errors at every number would not have figured in the analysis at all. On the other hand, we counted the errors in a way that was, if anything, biased against the knower-levels/guessing explanation.

Discussion

Overall, our results strongly support the knower-levels explanation, and show where counting and estimation error fit into a knower-levels framework. All the children in Dataset 1 knew how to count to ten, but they still made guesses about the cardinal meanings of the number words. In doing so, they made sure only to exclude cardinalities whose names they knew. In other words, their guesses were lower-bounded by their knowledge of number words. This is consistent with the knower-levels explanation.

Performance errors, unlike educated guesses, are symmetrical. They are not lower-bounded by a person's knowledge, because they do not reflect ignorance. An adult might correctly produce a set of 99 items, but then fail to produce a set of 100, giving only 98 by mistake. Similarly, a child who knows the meanings of all of the number words up through “ten” might correctly produce a set of 4 items, but then fail to produce a set of 5, giving only 3 by mistake. The fact that children do not make this type of error shows that their incorrect responses are lower-bounded by the numbers they know.

We did find some evidence of counting or estimation errors, especially for the target number word “six” in Dataset 1. Because children in Dataset 1 were asked for “six” only after they had produced “five” correctly, it is not surprising that the distribution of these errors was different—closer to the target and more symmetrical—than the other distributions we studied. Requests for “ten” also garnered more responses of 9 and 11 than requests for other number words. The errors around “ten” are particularly visible because they sit outside the range where most of children's guesses seem to fall (i.e., at the lower numbers or 15.)

Had performance errors accounted for the bulk of the data, and had we found evidence for performance errors around low-number targets as well, we might conclude that children really do understand the cardinal meanings of number words as soon as they learn to count. However, because such errors accounted for only a minority of the data, and only appeared for high-number targets, they seem to support the knower-levels view that children use counting to produce only large cardinalities (five and above), once they have figured out how counting works (Carey, 2001, 2004; Wynn, 1990, 1992).

Our conclusions from this study are both theoretical and methodological. On the theoretical side, we conclude that the knower-levels framework captures a real phenomenon in development. Children do appear to learn the cardinal meanings of the low number words “one,” “two,” “three” and perhaps “four” one at a time, and in order. During this period, they really do not know what cardinalities the higher number words denote, except that they are different from the cardinalities denoted by the low number words. This finding lends validity to those studies that have used the knower-levels framework in the past.

There are undoubtedly cases where children produce a correct response by accident, or where they make mistakes producing a low number whose meaning they do know. We saw one such example in Dataset 2, where a child who knew “two” (and presumably also knew “one,” according to the knower-levels view) did give 1 item when asked for “three.” But such cases are rare. The preponderance of data support the view that there are two types of Give-N trial: Trials where the child knows the number and produces the set correctly (with a small chance of counting or estimation error if the number is large), and trials where the child doesn't know the number, and makes a guess. But her guess will be an educated one, in that she will not give a set of objects whose number she can identify.

This study points to the need for a formal description of ‘chance’ performance on this task, or any task for which many different responses are possible. To assume *a priori* that all responses are equally likely, as if children were plucking their answers from a deck of cards or spinning them on a roulette wheel, is likely to lead us astray. Certainly it makes valid quantitative analyses more difficult. To specify the ‘chance’ or base rate distribution for this task requires a formal model beyond the scope of this paper, but we refer the reader to Lee and Sarnecka (2008) for such a model.

Finally, these results suggest that future studies of number development in young children should include a knower-level diagnostic task such as Give-N or What's-On-This-Card, and should analyze results by knower-level controlling for age, as well as by age controlling for knower-level. There is great individual variation in the ages at which children progress through the knower-levels, and any particular developmental change could be related to age, knower-level, or both. Knower-levels give researchers a relatively precise way to describe where an individual child is in the number-development process. This will enable much more powerful analyses of developmental data, and much more valid comparisons of groups of children, both within and across studies.

Acknowledgments

This research was supported by NICHD Grant 00234342 to the first author. Collection of Dataset 1 was supported by NSF REC-0337055 to Elizabeth Spelke and Susan Carey. Collection of Dataset 2 was supported by a fellowship from the University of Michigan Culture and Cognition Program to the first author and by NICHD Grant HD36043 to Susan Gelman. Collection of Dataset 3 was supported by NIH Grants R01 HD038338 and 1T32MH067556-01A1 to Susan Carey.

References

- Carey S. Cognitive foundations of arithmetic: Evolution and ontogenesis. *Mind & Language*. 2001; 16:37–55.
- Carey S. Bootstrapping and the origins of concepts. *Daedalus*. 2004; 133:59–68.
- Condry KF, Spelke ES. The development of language and abstract concepts: The case of natural number. *Journal of Experimental Psychology - General*. 2008; 137(1):22–38. [PubMed: 18248127]
- Cordes S.; Gelman, R. The young numerical mind: When does it count?. In: Campbell, JID., editor. *Handbook of Mathematical Cognition*. Psychology Press; New York: 2005. p. 127-142.
- Cordes S, Gallistel CR, Gelman R. Variability signatures distinguish verbal from nonverbal counting for both large and small numbers. *Psychonomic Bulletin & Review*. 2001; 8:698–707. [PubMed: 11848588]
- Dehaene, S. *The number sense: How the mind creates mathematics*. Oxford University Press; New York: 1997.
- Frye D, Braisby N, Lowe J, Maroudas C, Nicholls J. Young Children's Understanding of Counting and Cardinality. *Child Development*. 1989; 60:1158–1171. [PubMed: 2805894]
- Fuson, KC. *Children's counting and concepts of number*. Springer-Verlag; New York: 1988.
- Gallistel, CR. *The organization of learning*. MIT Press; Cambridge, MA: 1990.

- Gelman, R. A rational-constructivist account of early learning about numbers and objects.. In: Medin, DL., editor. *The psychology of learning and motivation. Advances in research theory*. Academic Press; San Diego: 1993. p. 61-96.
- Gelman R, Meck E, Merkin S. Young children's numerical competence. *Cognitive Development*. 1986; 1:1–29.
- Greeno JG, Riley MS, Gelman R. Conceptual competence and children's counting. *Cognitive Psychology*. 1984; 16:94–143.
- Le Corre M, Carey S. One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition*. 2007; 105:395–438. [PubMed: 17208214]
- Le Corre M, Van de Walle G, Brannon EM, Carey S. Re-visiting the competence/performance debate in the acquisition of the counting principles. *Cognitive Psychology*. 2006; 52(2):130–169. [PubMed: 16364281]
- Lee, MD.; Sarnecka, BW. A model of knower-level behavior in number concept development. 2008. Manuscript in preparation. [Downloadable from <http://www.socsci.uci.edu/~mdlee/bgm>]
- Platt JR, Johnson DM. Localization of position within a homogenous behavior chain: Effects of error contingencies. *Learning and Motivation*. 1971; 2:386–414.
- Sarnecka BW, Carey S. How Counting Represents Number: What Children Must Learn and When They Learn It. *Cognition*. in press.
- Sarnecka BW, Gelman SA. Six does not just mean a lot: Preschoolers see number words as specific. *Cognition*. 2004; 92:329–352. [PubMed: 15019554]
- Sarnecka BW, Kamenskaya VG, Yamana Y, Ogura T, Yudovina JB. From grammatical number to exact numbers: Early meanings of “one,” “two,” and “three” in English, Russian, and Japanese. *Cognitive Psychology*. 2007; (55):136–168. [PubMed: 17070794]
- Schaeffer B, Eggleston VH, Scott JL. Number development in young children. *Cognitive Psychology*. 1974; 6:357–379.
- Whalen J, Gallistel CR, Gelman R. Nonverbal counting in humans: The psychophysics of number representation. *Psychological Science*. 1999; 10:130–137.
- Wynn K. Children's understanding of counting. *Cognition*. 1990; 36:155–193. [PubMed: 2225756]
- Wynn K. Children's acquisition of number words and the counting system. *Cognitive Psychology*. 1992; 24:220–251.

Figure 1. Illustration of response curves predicted by two types of performance error: counting error (top) and estimation error (bottom). To minimize visual clutter, figure shows target number words “three,” “six” and “ten” only.

Figure 2a.

Incorrect responses from Datasets 1, 2, and 3, pooled. Arrows indicate the number asked for. Responses of 15 indicate that the child gave all 15 items to the animal.

Figure 2b.

Incorrect responses from Dataset 1 (skilled counters) only. Arrows indicate the number asked for. Responses of 15 indicate that the child gave all 15 items to the animal.

Figure 3.

Illustration of the analysis asking whether children ever gave incorrect answers that fell below a number they knew. “P” stands for “perfect,” indicating that this graph shows data from the subset of children who performed perfectly on 4. (I.e., they always gave 4 when asked for “four,” and never gave 4 otherwise.) “A” stands for “asked,” indicating that the data are from trials asking for five items. This is merely an illustrative example; the analysis was repeated for all possible pairs of numbers in each of the three data sets.

Table 1

Summary of Give-N Errors

Target number	Total trials for that target number	Number (%) of trials garnering incorrect responses	Mean incorrect response	Is the mean incorrect response different from the target number?
Datasets 1, 2 and 3 (pooled)				
one	831	51 (6%)	7.94	$t = 8.37, p < .001$
two	803	206 (26%)	4.37	$t = 8.05, p < .001$
three	870	441 (51%)	3.93	$t = 5.21, p < .001$
four	341	188 (55%)	4.75	$t = 2.72, p < .01$
five	823	621 (75%)	5.42	$t = 2.32, p < .05$
six	534	447 (84%)	4.72	$t = 6.97, p < .001$
eight	(all trials occurred in Dataset 1; see below.)			
ten	265	223 (84%)	5.50	$t = 15.41, p < .001$
Dataset 1 Only				
one	130	10 (8%)	11.60	$t = 6.28, p < .001$
two	102	16 (16%)	9.07	$t = 5.32, p < .001$
three	169	54 (32%)	6.81	$t = 5.32, p < .001$
four	126	60 (48%)	8.07	$t = 6.87, p < .001$
five	153	75 (49%)	8.51	$t = 6.57, p < .001$
six	74	24 (32%)	6.54	$t = 1.17, p = .25, N.S.$
eight	52	39 (75%)	8.84	$t = 0.47, p = .64, N.S.$
ten	54	43 (80%)	8.72	$t = 1.99, p = .10, N.S.$