Developmental Ordering, Scale Types, and Strong Inference

James A. Dixon
Trinity University

Developmental ordering, 1 item preceding another in development, is a primary piece of evidence in developmental research. However, testing developmental ordering hypotheses is remarkably difficult; the problem is that researchers rarely have absolute measures (ratio scales). Therefore, directly comparing 2 measures is often not sensible. This article demonstrates that, depending on the type of scale, the observed data pattern is constrained by the underlying relationship. Although the observed data pattern may not reflect the exact relationship, it does limit the possible relationships. Researchers can use this information to reject potential underlying relationships. An application of the method of strong inference (J. R. Platt, 1964) is outlined whereby researchers can use observed data patterns to systematically reject competing developmental ordering hypotheses. The approach requires only standard statistical methods and assumptions.

Developmental ordering is, arguably, the most important piece of evidence in developmental research. When one item precedes another in development (i.e., developmental ordering), researchers have a phenomenon that must be explained and an important constraint on developmental theory. Specifically, the developmental ordering of two items allows us to test whether one item is necessary for the development of the other. For example, do children need a certain degree of language ability to develop a theory of mind (Jenkins & Astington, 1996)? This hypothesis can be supported or refuted by examining the order in which children acquire the two items (e.g., specified language ability and theory of mind). Knowing whether a certain degree of language ability must develop first gives researchers important information about how children acquire a theory of mind.

Likewise, researchers can use the ordering of two items to test hypotheses about whether a common underlying structure explains the development of two items (e.g., Case, 1991; Griffin, 1991; O’Reilly, 1995). For example, Griffin proposed that children’s ability to interrelate dimensions, such as events and intentions, explains the development of many aspects of intrapersonal reasoning, such as attributing emotional states and moral reasoning. Abilities based on the same underlying structure should show developmental synchrony as that structure develops.

Unfortunately, testing developmental ordering hypotheses is remarkably difficult. The problem is a basic measurement issue rather than a statistical issue; simply, it is almost never clear that the measures of the two skills are comparable (Chapman & Chapman, 1973, 1978). To directly compare the measure of Skill A with the measure of Skill B, the researcher must know how the measure of each skill relates to the underlying level of the skill. Figure 1 shows the situation graphically. The true relationship between Skill A and Skill B is shown across development. In this example, children first acquire Skill A and then Skill B. Of course, researchers do not directly observe the true developmental relationship. Instead, the measures of each skill are compared with one another in hopes of uncovering the underlying developmental relationship.

Each underlying skill is represented by a vertical bar toward the center of the figure. The darker the area on each bar, the more advanced the developmental level. The rectangles, which surround a portion of each bar, mark the developmental period of that skill. That is, development begins at the bottom of the rectangle and is completed at the top of the rectangle.

As the child develops, he or she moves up the developmental timeline in the middle of the figure. The child’s level on each skill can be seen by placing a horizontal line across the developmental timeline. For example, the lowest dashed line shows a child early in development who has just started to acquire Skill A but has not yet begun to acquire Skill B.

On either side of each underlying skill is a vertical scale representing the observable measure of that skill. A hypothetical mapping between the underlying skill and the measure is shown by the lines connecting the underlying skill and the measure. The child’s observed performance can be seen by following the line connecting the underlying skill to the measure. The measures of each skill are the only observables in the figure. Therefore, it may be helpful for the reader to consider the figure as three-dimensional, with the measures coming off the page in depth. This emphasizes that development affects the measures only through changes in the underlying skill. The conclusions that researchers can draw about developmental ordering depend on how the measures are related to the underlying skills.

The purpose of this article is to show, with fundamental concepts from measurement theory (Coombs, Dawes, & Tversky, 1970; Stevens, 1951) and the method of strong inference (Platt, 1964), that (a) the patterns of data that researchers usually take as implying developmental ordering cannot be interpreted...
The acquisition of two skills, A and B, is shown across development. Each skill is represented by a vertical bar. The darker the bar, the more advanced the developmental level of that skill. All the developmental changes for each skill take place within the area marked by the rectangle. The measure of each skill is also shown. Because the measure is the only observable aspect of the figure, it may be helpful to view the measures as existing on the front plane of a cube and the other aspects as existing on the back plane. The mapping between the underlying skill and its measure is shown by the lines connecting them. The relationship between the underlying skills and the respective measures is ordinal in this example. Development of the underlying skill results in an increase in the observed measure. However, equal amounts of development in a skill do not always result in equal increases in the observed measure. Therefore, the scales in the figure are ordinal rather than interval.

Starting Assumptions

Scale Types

In developmental psychology and related fields, researchers most often have nominal, ordinal, or interval scale data (Achenbach, 1978; Stevens, 1951). Rarely do we have ratio scales (where zero on the scale corresponds to zero amount of the variable, and intervals are equal). Nominal data are not used to infer developmental ordering, for obvious reasons. If the possible values of the variable have no implicit order, then the variable...
DEVELOPMENTAL ORDERING 133

Figure 2. The data pattern usually interpreted as evidence that Skill A develops prior to Skill B. Skill A is shown on the vertical axis, and Skill B is shown on the horizontal axis. The defining characteristic of this data pattern is that it is curvilinear and opening downward.

cannot be used to say anything about developmental ordering. Therefore, the discussion here deals with ordinal and interval data.

Measures Are Valid

Researchers attempt to devise measures that capture all the developmental changes of interest in an item and to capture only those changes. If a measure does not capture all the developmental changes of interest, then some changes may occur that researchers cannot detect. This would limit our confidence in any conclusions about developmental ordering. Likewise, if our measures assess developmental changes that are not of interest, then differences in the observed measure cannot be confidently attributed to the item of interest. Again, this would limit the confidence researchers have in their conclusions. The logic presented here assumes that the measures assess only the item of interest and that all the developmental changes of interest can be detected. It is worth noting that the standard comparisons researchers routinely make, such as interpreting the difference between two means on a single measure, are problematic unless the measures are valid (Coombs, Dawes, & Tversky, 1970). Testing developmental ordering does not require assumptions about validity beyond those usually made (for a discussion of issues of validity, see Campbell & Stanley, 1963; Cook & Campbell, 1979; Krathwohl, 1985).

Cross-Sectional Design

The examples and logic presented below assume the researcher is using a cross-sectional design. Note that the measurement problem is not alleviated in any way by using a longitudinal design. Whether a researcher is tracking individuals across time or examining individuals of different ages, comparing two measures without knowing how they relate to the underlying variables of interest carries the same risk. For the sake of clarity, and because they are used so widely, the article focuses on cross-sectional designs. Each individual’s performance on two measures is assessed once, and both measures are taken simultaneously (or as close to simultaneously as reasonably possible).

Two Types of Evidence About Developmental Ordering

Currently, developmental researchers conclude that one skill develops prior to another skill (or that the skills develop synchronously) from two different types of evidence. First, researchers draw conclusions about developmental ordering when children from a single age group perform better on one skill than on another. For example, Xu and Carey (1996) showed that 10-month-old infants successfully used spatiotemporal information, or characteristics of objects in general. These same infants failed to use property-kind information, or specific features of the particular objects. They interpreted this interesting result as evidence that infants first form a superordinate “object” category and only later form categories for specific “objects.”

Second, researchers examine the pattern of children’s scores on the two items across the developmental span of interest and, depending on the observed data pattern, draw conclusions about the developmental ordering of the items (e.g., Case & Okamoto, 1996; Dixon & Moore, 1996; Jenkins & Astington, 1996; Moore, Dixon, & Haines, 1991; Wohlwill, 1973). For example, Figure 2 shows the observed relationship between two hypothetical skills, A and B. Each point in the scatter plot represents the performance of a different participant. This type of pattern is often interpreted as implying that Skill A develops prior to Skill B. The reasoning behind this interpretation is described in the following. Each child is caught at a particular point in develop-
ment. The least developmentally advanced children score poorly on both measures. The performance of these children is represented in the lower left corner near the origin. Somewhat more developmentally advanced children score well on Skill A, but very poorly on Skill B. Children who are yet more developmentally advanced score well on Skill A and moderately well on Skill B. At the highest developmental level, children score well on both measures. Interpreting the pattern of children’s scores on two or more measures has a long and venerable history in developmental psychology (e.g., Flavell, 1971; Froman & Hubert, 1980; Wohlwill, 1973) and in scaling theory (e.g., Coombs, 1964; Guttman, 1944).

Comparing Performance Within a Single Age Group

The first type of evidence, comparing the performance of a single age group on two measures, is very problematic. An example should clarify this difficulty. Ruffman & Keenan (1996) were interested in whether children, to understand the emotion of surprise, must first understand that people can hold false beliefs. They reasoned that because surprise occurs when a false belief is disconfirmed, an understanding of surprise would first require children to understand that people can believe something that is false. In their first experiment, they showed children two labeled and easily identifiable boxes (e.g., a tissue box and a light bulb box). Both boxes contained the same item (e.g., a light bulb) that was consistent with the label on one box but not on the other. Therefore, one box conformed to expectations, and the other did not. Children were shown the contents of both boxes. They were then asked what a doll would think was in each box. (This question, when asked about the box with the unexpected contents, is the belief question; it assesses whether children understand that the doll would have a false belief.) Children were also asked which box would make the doll feel surprised. (This is called the surprise question; the doll should feel surprised that the tissue box contains a light bulb.) Each child went through the procedure twice, each time with different types of boxes that had different contents.

Ruffman and Keenan (1996) found that, with only one exception, all the 5- and 6-year-olds in their study successfully answered both belief questions. However, over 30% of these same children answered one or more of the two surprise questions incorrectly. They concluded that understanding of false belief develops prior to understanding of surprise.

At first glance, this seems like a sensible conclusion. A sizable percentage of children did not correctly answer the surprise question but almost all correctly answered the belief question. The problem, of course, is that researchers do not know that the belief question measures the child’s understanding of false belief in the same way that the surprise question measures understanding of surprise. For example, what if the surprise question is simply more difficult than the belief question? The phrasing of the surprise question might simply make it more difficult to understand or to answer correctly. Or, the belief question might be a more sensitive measure of understanding false belief than the surprise question is of understanding surprise. The point here is not to advance either of these explanations of Ruffman and Keenan’s (1996) findings. Rather, the example demonstrates that because these alternative explanations are exceedingly difficult to eliminate, directly comparing the developmental level indicated by any two measures is problematic. Unless it is possible to tie the two measures being compared directly to the underlying variables (i.e., create ratio scales), it will never be possible to draw firm conclusions about developmental priority on the basis of the performance of a single age group.1

Interpreting the Pattern Across Development

The second type of evidence, the pattern of scores on the two items across the developmental span, cannot be literally interpreted either. Consider again the pattern shown in Figure 2. Does the shape of the observed data pattern, which seems to indicate that Skill A develops before Skill B, necessarily reflect the relationship between the underlying variables? And if not, can anything be concluded from the observed data pattern? As will be demonstrated in the next section, the answer to the first question will usually be “no”; the observed relationship is not necessarily the underlying relationship. But the answer to the second question is “yes”; interesting conclusions are still possible. The observed data pattern is constrained by the underlying relationship. Therefore, researchers can draw informative conclusions from the observed data patterns. The extent to which the observed pattern is constrained depends on the scale type.

How the researcher defines developmental priority and synchrony will also be very important. Both of these points are discussed in the next section.

Disconfirming Developmental Ordering Hypotheses

As Flavell (1971) noted, developmental ordering (i.e., priority and synchrony) can be defined in many ways. For example, one might define the developmental synchrony of two skills as (a) the two skills starting development at the same time or (b) the two skills starting development at the same time or (b)

1 The validity of the measures is a separate issue from their comparability. It is possible to have valid measures that cannot be directly compared because of differences in sensitivity, and so on. However, when the underlying variable is truly dichotomous rather than continuous, the distinction between validity and comparability blurs. With dichotomous variables, lack of sensitivity would mean that the change from “absent” to “present” was not detected. Therefore, an insensitive measure of a dichotomous variable will not be valid.
the two skills developing at the same rate. Depending on the research question, developmentalists use different definitions, either implicitly or explicitly. Of course, if one is interested in testing hypotheses about developmental ordering, clear definitions of the ordering hypotheses are necessary.

Figure 3 shows three important developmental ordering hypotheses. The top of the figure shows complete priority of one skill over the other. On the left side, Skill A completes development before Skill B starts development. Each skill is represented by a vertical bar. Darker points on the bar indicate higher levels of development, with the solid black bar indicating completed development of the skill. The right side shows the reverse ordering; Skill B completes development before Skill A begins to develop. Hypotheses that state that one skill must be in place before another can develop predict complete priority.

The middle of Figure 3 shows partial priority. The left side shows Skill A starting development before Skill B, but Skill B begins development before Skill A is complete. The right side shows the reverse situation. Hypotheses that state that one skill begins development before another predict partial priority.

The bottom of Figure 3 shows developmental synchrony. Skills A and B start and complete development at the same time. Hypotheses that state that the development of both skills is caused by a single underlying factor predict developmental synchrony, at least under some conditions.

Obviously, this is not a comprehensive list of developmental ordering hypotheses. However, the hypotheses listed above are frequently addressed in the field and are fundamental to building developmental theories. The approach discussed here could be extended to test other developmental ordering hypotheses.

**Ordinal Scale Data**

Suppose a researcher observed the relationship shown in the top of Figure 4. This relationship, as mentioned previously, is usually interpreted as developmental priority of Skill A over Skill B. Its identifying characteristic is that it is curvilinear and opening downward. Assuming the researcher only has ordinal scale data, what conclusions can be drawn from observing this relationship?

Figure 4 shows the two complete priority hypotheses in the second row. On each side, the relationship between the underlying skill and the measure is ordinal. The relationship between the underlying skill and the measure is depicted by the lines connecting them. Notice that development of the underlying skill results in some increase in the observed measure, but equal amounts of change in the skill do not always result in equal amounts of change in the measure. Therefore, the measure is only ordinal related to the underlying skill.

The left side of the second row of Figure 4 shows the underlying Skill A completing development before Skill B begins development. An observed pattern like that in the top of the figure can be produced by this relationship. This can be easily demonstrated by plotting the values of the observed measures at various points in development. Each dashed line indicates the level of a hypothetical participant on both skills, as he or she develops.

Because we observe the measure, not the skill itself, performance on each measure (found by following the connection from the skill to the measure) should be plotted. Numbers have been placed on the measure for convenience but are, of course, arbitrary.

The right side of the second row in Figure 4 shows the opposite ordering: Skill B completes development before Skill A. Notice that regardless of the ordinal mapping between each skill and measure, it is not possible to produce the pattern shown in the top of the figure. That is, the relationship between each skill and its measure could be exceedingly nonlinear (i.e., have very unequal intervals), but, as long as the ordinal relationship holds, the observed pattern cannot be produced. The reason for this is quite simple. If Skill B begins and completes development before Skill A begins to develop, then all the changes in Skill B must take place at the lowest observed level of Skill A. The observed data pattern shows that the changes in Skill B occur across many different levels of Skill A, which disconfirms this hypothesis. Therefore, if one has an ordinal scale or better, observing this pattern (curvilinear, opening downward) allows one to reject this developmental ordering hypothesis. This underlying relationship (Skill B starts and completes development before Skill A) cannot be responsible for what is observed.

The third row in Figure 4 shows the two partial priority hypotheses. On the left side, Skill A begins development before Skill B. This developmental ordering obviously can produce an observed pattern similar to that in the top of the figure. The right side of the third row shows the opposite partial priority hypothesis; Skill B begins development before Skill A. Surprisingly, when the scales are ordinal, this developmental ordering also can produce a data pattern like that in the top of the figure. For example, as shown on the right side, Skill B begins development first, but the early changes in the underlying skill produce very small changes in the measure. Later changes in Skill B produce much larger changes in the measure. Skill A shows the exact opposite pattern: early development produces large changes, and later development produces small changes. This situation will produce the observed data pattern. Therefore, when this data pattern is observed with ordinal scales, it is not possible to eliminate the opposite partial ordering hypothesis.

Likewise, an underlying relationship of developmental synchrony, shown in the bottom of Figure 4, can produce patterns like the one in the top of the figure. Observing such a pattern of data, therefore, does not eliminate the developmental synchrony hypothesis either. Despite the apparent inconsistency between the observed data pattern and these underlying relationships (i.e., synchrony or partial priority of B over A), it is possible for either of these relationships to produce the observed pattern. Therefore, they remain viable hypotheses. However, the hypothe-

Figure 4 (opposite). With ordinal scale data, observing the data pattern in the top panel, which is usually interpreted as Skill A developing prior to Skill B, allows researchers to reject the hypothesis that Skill B completes development before Skill A. All the other hypotheses in the figure remain viable.
sis that B starts and completes development before A cannot produce this pattern and can be confidently rejected.

The top of Figure 5 shows the data pattern usually interpreted as developmental synchrony. The characteristic feature here is that the two observed measures have a linear relationship. The relationship identified as synchrony (fourth row of Figure 5) obviously is consistent with the observed relationship in the top portion.

Neither of the complete priority hypotheses, shown in the second row of Figure 5, can produce this observed pattern, regardless of the ordinal mapping. The entire range of scores for the skill that develops first will be seen at the lowest level of the skill that develops second. Therefore, when this pattern is observed, both the complete priority hypotheses can be eliminated.

However, either of the partial priority relationships (third row of Figure 5) can explain the observed pattern. Depending on the mappings between the respective underlying skills and their measures, it is possible to observe a linear relationship, despite the asynchrony of the underlying skills. Both the examples of partial priority in Figure 5 will produce the synchrony data pattern because of the nonlinear mapping between the underlying skill and the measure. Therefore, when ordinal scales are used, observing a linear relationship between the measures does not eliminate either partial priority hypothesis. In summary, observing a linear relationship between the measures eliminates both complete priority hypotheses but is consistent with synchrony and both partial priority hypotheses when the scales are ordinal.

Interval Scale Data

With interval scale data, additional underlying developmental relationships can be eliminated when either data pattern is observed. Examine the second, third, and fourth rows of Figure 6, which show the developmental ordering hypotheses. The relationship between each underlying skill and its measure is shown by the lines connecting them. For each individual skill, equal amounts of development in the skill always yield equal amounts of change in the observed measure. The measures are, therefore, interval scales. This relationship between the underlying skills and the measures greatly constrains which hypotheses can explain the observed data pattern.

The top of Figure 6 shows the now familiar data pattern usually interpreted as Skill A developing prior to Skill B. When this priority pattern is observed with interval scale data, both the complete and partial priority hypotheses that specify B developing before A (right side of the second and third rows of Figure 6), can be rejected. Neither of these relationships can produce the observed pattern. The logic behind this conclusion requires careful attention to three relationships in the figure: the relationship between the two underlying skills across development and the relationship between each underlying skill and its respective measure. Because each underlying skill is proportionally related to its respective measure (i.e., interval scales), movement up the developmental time line results in changes in both measures that are proportional to the changes in the underlying skills. The consequence of the proportional relationship between the measures and the underlying skills is that the overall shape of the relationship between the two underlying skills across development will be reflected in the observed relationship between the measures. (The observed relationship may shift up or down or span smaller or greater regions of the measure, but its overall shape cannot change.) Therefore, because both the partial and complete priority hypotheses specify that a curvilinear, upward-opening data pattern should be obtained, the observed data pattern allows us to confidently reject these hypotheses.

Following the same logic, the developmental synchrony hypothesis (bottom row of Figure 6) can also be confidently rejected. Again, the overall shape of the relationship between the two underlying skills must be preserved in the observed pattern. The developmental synchrony hypothesis predicts that the relationship between the observed measures will be linear. Because the observed pattern violates this prediction, the synchrony hypothesis can be rejected.

When the pattern usually interpreted as developmental synchrony (top of Figure 7) is observed with interval data, the situation is even more clear. The other alternative hypotheses, both sets of complete and partial priority hypotheses, can be eliminated. Of the developmental relationships shown in Figure 7, only developmental synchrony can explain the observed linear data pattern. Both sets of complete and partial ordering hypotheses predict curvilinear patterns and, therefore, can be rejected.

With interval data, therefore, observing a “priority” data pattern eliminates both opposite priority relationships, complete and partial. The synchrony relationship is also eliminated. When the “synchrony” data pattern is observed, both the complete and partial priority hypotheses are eliminated. Data patterns with interval scales will eliminate more underlying relationships than those with ordinal scales, although ordinal scales still allow for some strong conclusions.

Significance Testing

Of course, it is important to test whether the observed pattern is statistically reliable. Standard methods, such as multiple regression and correlation (Cohen & Cohen, 1983) can be used to test whether the observed pattern is curvilinear, that is, has a significant quadratic component. These methods can also be used to argue that the relationship is linear, given sufficient power (Cohen, 1977; see Gigerenzer, 1993, for an alternative to standard significance testing). With these methods, it is possible to argue strongly about which data pattern has been observed. The logic outlined above can be applied, in conjunction with
standard statistical tests, to reject developmental ordering hypotheses.

Establishing Scale Types

It is important to recognize that establishing the scale type requires determining how the observed measure is related to the underlying variable. Without empirical evidence, one cannot assume that certain measures are interval scales. In fact, some researchers have pointed out that even ordinal scales must be empirically established (Cliff, 1993; Torgerson, 1958). Because interval scales are both more useful and more difficult to establish, I briefly outline below how empirical evidence for interval scales can be obtained (see Cliff, 1993, for a more thorough, but very accessible, discussion of establishing both interval and ordinal scales).

Interval scales are established by any one of a variety of methods. All of these methods might be described as “bootstrapping,” in that potential scale types are validated within the context of likely models. This may not be surprising because we are trying to establish the relationship between our measure and something that cannot be observed, the underlying variable. In each method, the relationships among three measures, (the measure of interest and at least two other measures), are empirically examined. The researcher also proposes a model that specifies how the underlying variables are related. In general, to the extent that the model fits the relationships among the observed measures, the scale type and the model are jointly confirmed (Anderson, 1970). The argument for this conclusion is that, other than the specified model and scale type, only a very limited number of models, combined with fairly unlikely relationships between the measure and the underlying variable, could produce the observed pattern of results. The specified model and scale type are accepted rather than the unlikely alternatives.

A brief example of one such “bootstrapping method,” functional measurement (Anderson, 1981, 1982), may clarify some details of the general approach. Schlottmann and Anderson (1994) investigated children’s understanding of expected value. Children were shown roulette-style spinners that were divided into red and blue sections. If the pointer landed on the red section, a fictional character won the specified prize. The size of the winning red section (probability) and the value of the prize (value) were independently manipulated. Children were asked to rate how happy the character would be (expected value) with each probability–value combination. Schlottmann and Anderson predicted that older children should follow the normative multiplicative rule:

\[
\text{Expected Value} = \text{Probability} \times \text{Value}
\]

Younger children were expected to follow an additive rule. As predicted, the younger children (5 and 6 years) showed an additive pattern, and older children (8 years and above) showed a multiplicative pattern. Schlottmann and Anderson (1994) noted that two interpretations of the developmental shift from an additive to a multiplicative pattern are possible. First, the ratings are an interval scale, and there is a developmental shift from children’s using an additive to a multiplicative rule for judging expected value. Second, children across all ages in the study used the same rule (i.e., additive or multiplicative), but the developmental shift in the judgment pattern results from developmental differences in how the judgment scale was used (either a shift from linear to logarithmic or antilogarithmic to linear, depending on whether an additive or a multiplicative model is proposed).

Schlottmann and Anderson (1994) argued convincingly against the second interpretation, citing evidence from a number of studies that do not show a developmental shift in how the response scale is used. Therefore, they concluded that the observed ratings were linearly related to the underlying subjective impressions of expected value; that is, the measure is an interval scale. Having established an interval scale for the subjective impression of expected value, the researchers would be in a very powerful position to test hypotheses about its developmental precursors and successors (for a more thorough discussion of the issues involved in applications of the functional measurement approach, see Anderson, 1976, 1981, 1982; Surber, 1984).

Establishing interval scales is a somewhat arduous task. However, for researchers interested in testing hypotheses about developmental priority, the enormous benefit, in terms of the allowable conclusions, will be well worth the effort.

Examples of Developmental Ordering Studies

I noted at the outset of the article that application of this logic does not require any assumptions beyond those made for standard data analytic techniques. Specifically, the scales must be either ordinal or interval, the measures must be valid, the design must be cross-sectional (to simplify the discussion), and the measures must be taken simultaneously or nearly so. Given this set of easily met assumptions, a huge number of studies could reasonably apply the approach outlined here. Below, I briefly discuss two studies that could apply the approach either given their current data or after gathering further data. The studies were chosen because they explicitly discuss different developmental ordering hypotheses and represent quite different developmental paradigms.

Development of the Object Concept in Infancy

Xu and Carey (1996) presented evidence that infants use spatiotemporal information to establish the identity of objects before they use property-kind information. This developmental

---

Figure 6 (opposite). With interval scale data, observing the data pattern in the top panel, which is usually interpreted as Skill A developing prior to Skill B, allows researchers to reject both hypotheses that specify Skill B developing before Skill A (complete and partial priority). The synchrony hypothesis is also rejected. Both priority hypotheses, which specify that Skill A develops prior to Skill B (complete and partial), remain viable.
ordering has important implications for early conceptual development. A central piece of evidence in support of their hypothesis comes from comparing how long infants look at different displays after experiencing one of two conditions: property-kind or spatiotemporal (Experiment 2). In the property-kind condition, an infant is first presented with a solid screen. An object is brought out from behind the screen and displayed briefly to one side. The object is then returned to its place behind the screen. A second object, with an entirely different appearance, then is brought out from behind the screen and is briefly displayed before being returned to its place behind the screen. The screen is then lowered revealing either (a) the two objects or (b) only one of the two objects. In this condition, the two objects are never presented at the same time. Therefore, only if the infants notice that the two objects are different, on the basis of their different properties, should the single object be surprising and looked at longer.

The spatiotemporal condition is identical to the property-kind condition, with one exception: Infants see both objects brought out simultaneously from behind the screen, just after the screen is first presented. The rest of the sequence is exactly the same, including the screen eventually being lowered to reveal both objects or only one object. In this condition, infants have spatiotemporal information in that both objects are presented at one time (i.e., they are separated spatially at a given moment in time). Therefore, when a single object is presented, infants should look longer at it.

In the property-kind condition, the difference in looking time between the presentation of a single object (the unexpected event) and that of the two objects (the expected event) measures the infants' use of property-kind information. In the spatiotemporal condition, the difference in looking time between one and two objects measures the infants' use of spatiotemporal information in addition to property-kind information.

Xu and Carey (1996) hypothesized that infants first develop the ability to use spatiotemporal information and only later develop the ability to use property-kind information. One interpretation of this hypothesis is that it specifies complete priority. That is, children's development of the ability to use spatiotemporal information is complete before their ability to use property-kind information begins to develop. One prediction that can be drawn from this hypothesis is that if infants who are at the appropriate point in the developmental sequence are sampled, the difference in looking time should be greater for the spatiotemporal condition than for the property-kind condition. In addition to having to sample infants at a specific point in the developmental sequence, this prediction also requires that the two measures be directly comparable (comprise ratio scales). For example, the spatiotemporal and property-kind information must be equally salient, otherwise a difference between the two measures will not be easily interpretable.

Another way to address the developmental ordering of these two items is to examine the relationship between the two measures across development and apply the logic presented above. In addition to the 10- and 12-month-olds tested by Xu and Carey (1996), it would be most effective to sample infants from across the entire developmental range for these two items. The two measures would then be cast in a two-dimensional scatter plot, as in Figure 2, with each infant providing a single data point. Assuming that the measures (differences in looking times) are ordinally related to the underlying variables, observing a linear relationship between the two measures, as in the top of Figure 5, would disconfirm Xu and Carey's hypothesis as well as the opposite complete priority hypothesis. Likewise, observing a curvilinear relationship consistent with their hypothesis, as in the top of Figure 4, would also disconfirm the opposite complete priority hypothesis. Disconfirming the partial priority and the synchrony hypotheses would require interval scale measures.

Development of Central Conceptual Structures for Number

Case and Okamoto (1996) proposed that a set of central conceptual structures are used for reasoning about the domain of number. In general, these underlying structures gradually develop and eventually merge, such that children can simultaneously consider multiple dimensions at one time. Case and Okamoto presented a compelling model of how this development occurs. Their theory predicts an interesting developmental ordering. Because a single set of conceptual structures is used for the entire numerical domain, performance on diverse tasks that involve number should show developmental synchrony.

Case and Okamoto (1996) presented evidence that bears on this prediction (Studies 1 and 2). Children from ages 6 through 10 were presented with two sets of tasks. Each task was carefully designed to assess the child's current level of functioning in one area of number. One task assessed knowledge of number, the other assessed word problem skill. Both these areas should be directly related to the underlying conceptual structure for number and, therefore, develop synchronously.

The logic outlined here could be very effectively applied to this study. Case and Okamoto (1996) presented a scatter plot of scores on the "word problem" task as a function of scores on the "number knowledge" task. It is interesting that the two sets of scores were highly correlated (r = .79), and the plot appears quite linear (although a test of the quadratic component is not reported). If Case and Okamoto could demonstrate that, in addition to being linearly related, their measures of number...
knowledge and word problem skill were interval scales, they could persuasively eliminate the major competing hypotheses about developmental ordering. Both complete and partial priority hypotheses would be eliminated. This would be an exceptionally strong piece of evidence for Case and Okamoto's theory.

Conclusions

When the data are ordinal or interval, the data patterns that are usually interpreted as evidence of developmental ordering do not allow researchers to confirm the developmental ordering that seems apparent in the data pattern. That is, they cannot be interpreted literally. In all but one case (observing the synchrony pattern with interval scale data), more than one underlying relationship could be responsible for the observed pattern. However, each of the observed data patterns eliminates some underlying relationships. By taking advantage of the fact that the observed pattern could only be produced by a certain set of underlying relationships, researchers can strongly reject ordering hypotheses that are not in that set.

This approach allows researchers to draw conclusions about developmental ordering without assuming that the scales are comparable. Innumerable studies have assessed developmental ordering by comparing performance within a single age group or literally interpreting a pattern of scores across developmental levels. Almost without a doubt, the developmental orderings reported in many of these studies are correct. For example, Ruffman and Keenan's (1996) finding that children understand false belief before they can understand surprise seems entirely sensible. However, stronger evidence on developmental ordering is readily available. The approach outlined here offers a simple way for researchers to make stronger arguments about developmental ordering without making any additional assumptions.

This approach can be considered as an application of the method of strong inference (Platt, 1964). In strong inference, the researcher proposes a set of mutually exclusive hypotheses. In the current context, the hypotheses are about the developmental ordering of the two items in question. The observed data pattern disconfirms one or more of these hypotheses, but does not confirm any single hypothesis. Consistent with strong inference, scientific progress is marked by disconfirming hypotheses.

Another feature of strong inference is that, after eliminating an initial set of hypotheses, researchers recycle through the process in hopes of further restricting the number of possible explanations; this will be useful here as well. For example, suppose a researcher interested in testing among the synchrony and the two complete priority hypotheses initially observes the "priority" pattern, using ordinal scale measures. This result is very informative; it eliminates one of the two complete priority hypotheses. However, to distinguish between the remaining two hypotheses, the researcher must recycle through the process, using interval scale measures.

Depending on which developmental ordering hypotheses are being tested, the researcher can use ordinal measures, which are much easier to construct, or the researcher must go through the more arduous process of constructing measures that are interval scale. For example, if the researcher is interested in testing between the two complete priority hypotheses, then ordinal measures will suffice. However, if the researcher wants to test between partial priority and synchrony, ordinal scale measures will be of no use because any observed pattern will be consistent with all the hypotheses of interest. Interval scale measures will be required to distinguish between these hypotheses.

The developmental hypotheses discussed here explicitly specify a single developmental pathway. That is, all children are hypothesized to follow a single developmental sequence. These types of hypotheses are commonly proposed and evaluated. In principle, multiple pathway hypotheses could also be tested in the way described here, but hypotheses about single developmental pathways will be easiest to test. Many data patterns will be consistent with multiple pathway hypotheses.

Likewise, the observed data patterns that we have considered seem to suggest a single developmental pathway. These data patterns are especially relevant because they are often interpreted as confirming developmental orderings, but other patterns will also disconfirm ordering hypotheses. The critical question will be whether the observed pattern could have been produced by the hypothesized relationship. If not, then the hypothesized relationship can be rejected.

Developmental ordering has been, and will continue to be, a very important source of evidence for developmental research. The orderings of items as they emerge in development are phenomena that must be explained and that constrain our theories. For example, research with infants continues to show that very young children have surprisingly sophisticated abilities (e.g., Dannemiler & Stephens, 1988; Katz, Cohn, & Moore, 1996; Spelke, 1988). Part of what makes these findings so interesting is that we did not expect to find these abilities at this point in the developmental ordering. The fact that an infant has a fairly sophisticated understanding of the physical world (Spelke, 1988) strongly impacted developmental theory because many believed that children did not yet have the concepts necessary to understand the physical world (e.g., object concept).

There are many different ways to define developmental ordering. The approach presented here requires researchers to think explicitly about which developmental ordering hypotheses that they are testing. Proposing and testing specific hypotheses will advance our understanding of development more quickly. With the approach outlined here, researchers can test specific developmental ordering hypotheses, using the data and statistical methods they have in hand.

References


Received January 13, 1997
Revision received June 26, 1997
Accepted June 26, 1997