JOURNAL OF APPLIED MEASUREMENT, *15*(4),

Does the Model of Hierarchical Complexity Produce Significant Gaps between Orders and Are the Orders Equally Spaced?

Michael Lamport Commons
Harvard Medical School

Eva Yujia Li
Harvard Graduate School of Education

Andrew Michael Richardson
Salem State University

Robin Gane-McCalla
Dare Institute

Cory David Barker
Antioch University Midwest

Charu Tara Tuladhar
Mount Holyoke College

The model of hierarchical complexity (MHC) provides an analytic a priori measurement of the difficulty of tasks. As part of the theory of measurement in mathematical psychology, the model of hierarchical complexity (Commons and Pekker, 2008) defines a new kind of scale. It is important to note that the orders of hierarchical complexity of tasks are postulated to form an ordinal scale. A formal definition of the model of hierarchical complexity is presented along with the descriptions of its five axioms that help determine how the model of hierarchical complexity orders actions to form a hierarchy. The fourth and the fifth axioms are of particular importance in establishing that the orders of hierarchical complexity form an equally spaced ordinal scale. Previously, it was shown that Rasch-scaled items followed the same sequence as their orders of hierarchical complexity. Here, it is shown that the gaps between the highest Rasch scaled item scores at a lower order and the lowest scores at the next higher order exist. We found there was no overlap between the Rasch-scaled item scores at one order of complexity, and those of the adjoining orders. There are "gaps" between the stages of performance on those items. Second, we tested for equal spacing between the orders of hierarchical complexity. We found that the orders of hierarchical complexity were equally spaced. To deviate significantly from the data, the orders had to deviate from linearity by over .25 of an order. This would appear to be an empirical and mathematical confirmation for the equally spaced stages of development.

Requests for reprints should be sent to Michael Lamport Commons, Harvard Medical School, 234 Huron Avenue, Cambridge, MA 02138, USA, e-mail: commons@tiac.net.

The model of hierarchical complexity is a framework to represent the intelligence of humans and animals and provides an alternative to the well-known IQ test. In this framework, intelligence is measured by the complexity of tasks that an individual accomplishes. The model of hierarchical complexity has been constructed into a measurement system, where order of tasks is a scale that measures the complexity of tasks. When constructing a scale, it is always important to validate and understand the properties of the scale. This paper explores whether the orders of hierarchical complexity form an ordinal scale. If the orders of hierarchical complexity do form an ordinal scale, then there should be discernible gaps between each order. It also explores whether it is a linear and equally spaced scale. There are specific reasons for us to be concerned with these properties of the scale.

First of all, whether or not the order of hierarchical complexity is an ordinal scale is essential to the validity of this scale (Krantz, Luce, Suppes, and Tversky, 1971). If the order of hierarchical complexity is an ordinal scale, it means that the order is a valid representation of the complexity of items. Past research, which tested participants on items with different orders of hierarchical complexity, showed that items with higher orders were always more difficult than items with lower orders. This paper explores whether there are difficulty "gaps" between items of different orders of hierarchical complexity (Commons and Calnek, 1984).

Second, whether a scale is equally spaced has implications on what inference can be drawn from the scale. To have orders of hierarchical complexity as an equally spaced scale would mean that moving from one order to the next is always the same increase in difficulty. In addition, an equally spaced scale would indicate that orders of hierarchical complexity do not only represent the relative position of the difficulty of tasks, but also the quantity of the difficulty.

Introduction to model of hierarchical complexity

The model of hierarchical complexity (MHC) is a measurement theory which analyzes the difficulty of tasks, which is represented by the *orders of hierarchical complexity*. Model of hierarchical complexity is not the only theory of development based on task complexity. Other metrics of task complexity have been proposed as well. Horizontal, classical or traditional information complexity is one of them. It describes the number of "yes-no" questions (Krippendor, 2009; Shannon and Weaver, 1948). In classical information complexity, if a task requires one such question, the answer would consist of 1 bit of "horizontal" information. Similarly, if a task requires two such questions, the answers would transmit 2 bits. Each additional 1-bit question would add another bit. Horizontal complexity, then, is the sum of bits required by tasks that require "yes-no" questions. The total number of actions is 2^n, and the number of bits = n.

Older metrics of task complexity such as the horizontal complexity and others have a number of limitations. What is promising about the model of hierarchical complexity is that it is a newer model that overcomes those limitations. The MHC does not confound stages with amount of information. Model of hierarchical complexity is based on vertical complexity that involves hierarchical information. Hierarchical complexity refers to tasks hierarchy that requires the performance of lower-order tasks in order to perform more complex, higher order tasks.

Hierarchical complexity has several advantages over horizontal complexity. The advantages will be discussed in greater detail below.

Advantages of hierarchical complexity

Hierarchical complexity is better at explaining problem solving difficulty than horizontal complexity. Evidence for this claim comes from the study done by Commons (2008a). Participants were asked to solve four sets of problems from mathematics and science domains: beam balance (derived from Inhelder and Piaget, 1958; algebra; infinity; and laundry problems (derived from the pendulum problem of Inhelder and Piaget, 1958). A stepwise regression performed to test whether variables such as *number of calculations required*, *number size, place in order* and *hierarchical complexity* predicted performance on those four tasks. The number of calculations required yields the traditional measure of bits. The results showed that hierarchical complexity alone accounted for far and away the most variability in all the four tasks: beam balance ($\beta = 0.977$, $p < 0.001$), algebra ($\beta = 0.961$, $p < 0.001$), infinity ($\beta = 0.921$, $p < 0.001$) and laundry problem ($\beta = 0.964$, $p < 0.001$).

When other smaller variables, which include horizontal complexity, were added to the regression model along with hierarchical complexity, hierarchical complexity wiped out the effect of all those variables. Whereas most of those smaller variables did not have a significant effect on performance, even the ones that had significant effect did not have a large effect on performance. In the beam balance problem, calculation required (traditional bits) had a small but significant effect on performance ($\beta = 0.218$, $p < 0.001$). Adding the number of calculation required (traditional bits) to the model after hierarchical complexity, did not account for much more than the amount of variation accounted by hierarchical complexity alone, $\Delta R^2 = 0.025$. Similarly, in the algebra problem, effect of calculation required and hierarchical complexity combined was not much greater than the effect of hierarchical complexity alone, $\Delta R^2 = 0.013$. The effect of hierarchical complexity ($\beta = 0.943$, $p < 0.001$) was greater than that of calculations required ($\beta = 0.117$, $p < 0.024$) in this problem as well. For infinity problem, the effect of other variables was negligible and was not significant. Hierarchical complexity was the only variable that predicted performance in this problem ($\beta = 0.912$, $p < 0.001$). For laundry problem, the amount of variation in performance accounted for by hierarchical complexity and calculation required combined was not much greater than the amount of variation accounted for by hierarchical complexity alone, $\Delta R^2 = 0.010$. Calculations required only had a small but significant effect ($\beta = 0.168$, $p < 0.001$). Data from this study showed that hierarchical complexity accounted for variation in performance more than other horizontal complexities (bits) did. Hence, Hierarchical complexity is a better predictor of performance than other horizontal complexities.

Second advantage of hierarchical complexity is that it is a clear scoring scale and can be used to score narratives and vignettes as well. In most horizontal complexities there is no clear way of applying them to narratives and vignettes.

Third advantage of hierarchical complexity is that it captures difficulty in hierarchical nature. Hierarchical nature entails requirement of completion of other tasks before the completion of one task. This is left out in horizontal complexity. Hierarchical complexity contains horizontal complexity. The orders of hierarchical complexity are also related to 2^n, where n is the ***order or hierarchical complexity*** of the task. Horizontal complexity has the same formula. However, in hierarchical complexity there can be more actions that do not change order. It also integrates hierarchical relations among actions.

The fourth advantage of hierarchical complexity is that it does well in predicting performance in scientific and mathematical domains as well as social domains whereas, horizontal complexity cannot be used to predict performances in such domains as there is no way of coding number of actions involved.

Axioms and Definitions of model of hierarchical complexity

In the model of hierarchical complexity, successful completion of a task at a certain *order* of hierarchical complexity indicates the person or animal is performing at the *stage* that has the same number and name as that order of Complexity. The model has been broadly applied to constructing assessment tests in the field of stages of social perspective-taking, general logic, problem solving, etc. (Bernholt, Parchmann, and Commons, 2009; Commons, Goodheart, Pekker, Dawson, Draney, and Adams, 2008; Commons, Rodriguez, Adams, Goodheart, Gutheil and Cyr, 2006; Dawson, 2002, 2003; Skoe, (in press)

It is important to understand the concepts *of actions, events and tasks* to understand MHC. ***Actions*** are defined as behavioral events that produce outcomes. Actions may be attributed to organisms, social groups, and computers. Actions may be combined to produce new, more complex actions (Binder, 2000). *Events*, including behavioral events, are perturbations that can be detected by at least two independent paths (Commons, 2001). A *task* can be defined as a set of required actions that obtain an objective, though the performed actions may or may not complete a given task.

O***rder of hierarchical complexity*** characterizes the underlying difficulty of tasks. The higher the order of hierarchical complexity, the more the difficulty of the task is. A task analysis allows for specification of this order. Past research has defined 16 orders of hierarchical complexity, as shown in Table 1.

Table 1

16 Stages and orders of hierarchical complexity

Order	Name of Complexity
0	Calculatory
1	Sensory and Motor
2	Circular Sensory-Motor
3	Sensory-Motor
4	Nominal
5	Sentential
6	Preoperational
7	Primary
8	Concrete
9	Abstract
10	Formal
11	Systematic
12	Metasystematic
13	Paradigmatic
14	Crossparadigmatic
15	Meta-Crossparadigmatic

The most irreducible task is at order 0. Order 0 actions are not planned or controlled. Examples of order 0 tasks are computer running a written computer program. order 0 consists of traditional complexity that computers begin with, bits taking values of 0 and 1. All actions to complete the task are exact as written by the programmer. There is no flexibility. Order 1 actions are flexible and adaptive, as compared to order 0 actions. Some examples are tropism and bodily movements elicited by simple reflex. In these cases, the organisms respond to external stimuli. Order 2 actions are more complex. They are made out of coordinating order 1 actions. One example is a baby reaching out to breast when it is hungry. The tasks of order 3 are made out of actions of the order 2, and so on. The repeating process of an order of actions defined in terms of lower order actions produces the numerical relation structure, and stratifies orders of hierarchical complexity.

Higher order task action is: a) defined in terms of tasks at the *next lower* order of hierarchical complexity task action; b) defined as the higher order task action that organizes two or more less complex actions; that is, the more complex action specifies the way in which the less complex actions combine; c) defined as the lower order task actions have to be carried out *non - arbitrarily.* Once these conditions have been met, we say the higher order task *coordinates* the tasks of the next lower order.

For example, simple calculations of addition and multiplication are primary order 7 tasks (Commons, Miller, Goodheart, and Danaher-Gilpin, 2005). Multiplication is not hierarchically more complex than addition, because it does not fulfill all the conditions mentioned above. It fulfills condition (a): repetitive addition defines multiplication. However, it does not fulfill conditions (b) or (c) because the organization of addition is not non-arbitrary. It could occur in any order. In contrast, the task of calculating $a \times (b + c)$ is at the concrete order 8. It fulfills all three conditions listed above. First, calculating $a \times (b + c)$ is defined of both multiplication and addition. Second, it organizes the multiplication and addition in a non-arbitrary way. One has to do $(b + c)$ before doing $a \times (b + c)$. Or one has to do $(a \times b) + (a \times c)$ in the order specified by mathematical rules.

Figure 1 illustrates the hierarchical structure of tasks as described by the model. The measurement system of the model of hierarchical complexity is composed of axioms. Axioms are rules that are followed to determine how the model of hierarchical complexity orders actions to form a hierarchy. There are five axioms: *well ordered, transitive, chain rule, coordination rule* and *equal spacing* (optional). The concatenation

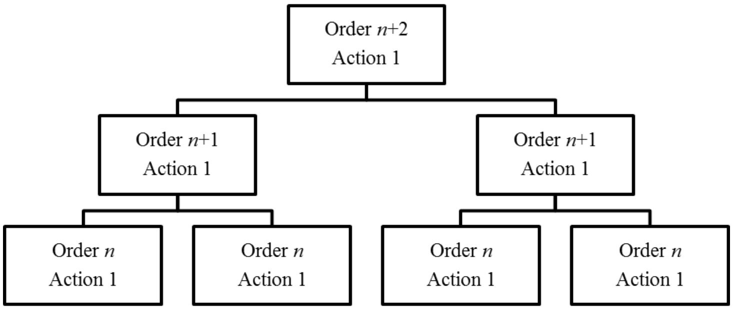

Figure 1. Order of hierarchical complexity

operator, "○", represents the way in which actions are connected. A system of entities, as a set of actions, is represented by letters such as *a*. The comparison operator, " ≻" is used to arrange actions in a hierarchy. In the case of real numbers, the comparison operator is ">" and the concatenation operator is "+". If *a* is an *n* order action the assignment function φ assigns the number *n* to *a* which is denoted by $\varphi(a) = n$. The assignment function, $\varphi(a)$, denotes the *order of hierarchical complexity* (OHC).

Axiom 1, *Well ordered*: If $a \succ$ b, then $\varphi(a) > \varphi(b)$

Axiom 1 means that when actions are converted to numbers by applying the mathematical assignment function φ, action *a* remains more hierarchically complex than action *b*.

Axiom 2, *Transitivity*: If $a \succ b$ and $b \succ c$ then $a \succ c$

Axiom 2 means that if action *a* is more complex than action *b,* and action *b* is more complex than action *c,* then action *a* is more complex than action *c.*

Axiom 3, *Chain rule:* $\varphi(a \circ b) = \max(\varphi(a), \varphi(b))$ if $\varphi(a \circ b) = \varphi(b \circ a)$

Axiom 3 states that when actions *a* and *b* are chained together in some order, and the order in which they are executed is not influential to accomplishing a task, the order of hierarchical complexity of $(a \circ b)$ equals that of the highest subaction. Chaining together the two actions does not produce an action that is hierarchically more complex than either of the subactions.

Axiom 4, *Coordination rule*: $\varphi(a \circ b) = \max(\varphi(a), \varphi(b)) + 1$ if $\varphi(b) = \varphi(a)$ and $\varphi(a \circ b) \neq \varphi(b \circ a)$.

In this case, the concatenation operator "○" coordinates the organization of the ordering of action rules in a non-arbitrary way. In addition, action *a* and action *b* has to be on the same stage. When these two conditions are satisfied, the coordination of action *a* and action *b*, which is represented by $(a \circ b)$, is one order more complex than either of the subactions. $\varphi(b) = \varphi(a)$ is necessary because, in order for the coordinated action to move up a stage, the actions have to be on the same stage. The coordination of two actions on different stages does not produce an action that is one stage higher.

Axiom 5, Optional, *Equal spacing* (optional): $\mathrm{OHC}(n + 1) - \mathrm{OHC}(n) = 1$

where, $\mathrm{OHC}(n) = \varphi(a)$, then for every order n, $(n)(\mathrm{OHC}(n + 1) – \mathrm{OHC}(n)) = 1$

Axiom 5 states that the a priori difficulty of a task action changes by 1 for each change in the order of hierarchical complexity, irrespective of what adjacent orders of Hierarchical Complexities one is comparing.

Axiom 4 and axiom 5 are of particular significance in establishing that there are significant gaps between orders and that the orders are equally spaced. The above five axioms allow more specific definitions about the model of hierarchical complexity which help establish the above argument.

Definition 1: There exists *simple actions*, x with $\varphi(x) = 1$. This is the lowest order action.

Definition 2: If there is no action, then the null action is at order 0.

Definition 3: A higher order hierarchically complex action is defined in terms of two or more next lower order actions, $A = \varphi(B \circ C)$, where $\varphi(B)$ and $\varphi(C)$ are both less hierarchically complex than $\varphi(A)$ if $\varphi(B \circ C)$ is a coordination. This creates the hierarchy:

$A = \{a, b\}$ a, b are lower order of hierarchically complexity than A and together composes set A

$A \neq \{A, ...\}$, According to the definition of a set, a set cannot contain itself. This can be seen from Russell's Paradox. That paradox points out the contradiction in the definition of a set. If a set is not a member of itself, it would qualify as a member of itself by the same definition (See Russell's paradox (1902; 1980).

This definition follows definition 1 and Axiom 4.

Next, the differences between chain rules and coordination rules are explained in more depth.

Definition 4: Given a permutation of concatenated actions = $(i_1, i_2, \ldots, i_n)$ of the natural numbers 1, 2,…, n, the execution of action A is simply $A_{i1} \circ A_{i2}, \ldots, \circ A_{in}$.

The rule, R, is a chain rule if the outcome of the action is the same for all $n!$ permutations of the numbers 1, 2,…, n. The outcome of the order of actions, $A_{i1} \circ A_{i2} \circ \ldots \circ A_{in}$ is the same for all permutations $(i_1, i_2, \ldots, i_n)$ of 1, 2,…, n.

Rule, R, is a coordination rule if there exists at least one permutation of actions $R = (j_1, j_2, \ldots, j_n)$ of the numbers 1, 2,…, n so that the execution of the actions A_i i.e., $A_{j1} \circ A_{j2}, \ldots, A_{jn}$, is *not* the same as the outcome of the action A. Hence, the outcome of A_i is given by at least one, but *not all*, permutations of the A_i. This extends similarly to the cases where A consists of infinitely many actions.

Note that by Axiom 4, a coordination action $A = (\{A_1, \ldots\}, R)$ necessarily coordinates subactions of subtasks of equal orders of hierarchical complexity (i.e., $\varphi(A_1) = \varphi(A_2) = \ldots$). Thus the order of hierarchical complexity of A is one higher than the order of hierarchical complexity of all its subactions. Therefore, A_1 may be replaced by any subaction of A and still obtain the same result. As a consequence of these axioms, we see that if we let $\boldsymbol{A}$ denote the collection of all actions in a given system, then the order of hierarchical complexity is a function h: $\boldsymbol{A} \rightarrow \boldsymbol{N}$, where $\boldsymbol{N} = \{0, 1, \ldots\}$ is the set of natural numbers (and zero) under the usual ordering.

The following properties emerge from the axioms and the definitions:

1. ***Discreteness***: The order of hierarchical complexity of any action is a nonnegative integer. In particular, there are gaps between orders.
2. ***Existence***: If there exists an action of order n and an action of order $n + 2$, then there necessarily exists an action of order $n + 1$.
3. ***Comparison***: For any two actions A and B, exactly one of the following holds: $\varphi(A) > \varphi(B)$, $\varphi(A) = \varphi(B)$, $\varphi(A) < \varphi(B)$. That is, the orders of hierarchical complexity of any two actions can be compared.
4. ***Non-reducibility***: A higher order action cannot be equal to any lower order actions. This property arises from the coordination rule, which claims that the coordination of two or more actions at the same order produces an action that is one order above.

Concepts from set theory are applied here to clarify why two order tasks can be non-arbitrarily ordered only at the next order. The higher order corresponds to a set A. Assume $A = \{a, b\}$. The lower order relations in the system correspond to the elements of lower order elements of the set, actions a and b. This creates the hierarchy because the set A is not the same as its elements a and b. The elements are at a lower order than the set. Therefore, the order of the set is not equal to the order of its elements, and $n + 1 \neq n$. Hence, the orders cannot be collapsed.

For example, consider an empty set $\underline{\emptyset}$. Russell argued that an empty set cannot be a member of itself (Godehard, 2004). An empty set $\underline{\emptyset} = \{\ \}$ has no member. Having no members mean that there is nothing in it, or the member is "nothing". Because Ø is a set, it is "something". Something cannot equal to nothing. Therefore, an empty set Ø cannot equal to its member. Likewise, a higher order action cannot equal to any lower order action from which it is made.

A way to avoid this paradox is Russell's type theory. First a hierarchy of types is created, and then each mathematical (and possibly other) entity is assigned to a type. Objects of a given type are built exclusively from objects of preceding types (those lower in the hierarchy) to preventing loops. The same is true for orders of hierarchical complexity

This is consistent with Inhelder and Piaget and the model of hierarchical complexity. These theories state that each next order actions coordinates the actions performed at the preceding order of complexity. To apply the premise successfully, the actions of each stage must be unambiguously specified. The stage generator concept successfully eliminates ambiguity about what makes a stage by precise specification. Given a collection of actions *A* and a participant *S* performing *A*, the *stage of performance* of *S* on *A* is the highest order of the actions in *A* completed successfully.

Stage(S, A) = max $\{h(A) \mid A \in A$ and A completed successfully by $S\}$.

Empirical Study

In past research, whether orders of hierarchical complexity of items truly account for difficulty of tasks has been empirically tested. In several studies (Commons, Goodheart et al. 2008; Bernholt, 2009), items were constructed using the Model of Hierarchical Complexities theory. Items at order (n +1) coordinated items at order n, and the organization was non-arbitrary. Participants were asked to complete these item tasks. Answers were marked right or wrong. Then the data were analyzed using Rasch analysis (Bradley and Terry, 1952; Luce, 1959, Rasch, 1980).

In Rasch analysis (1980), probability of items being answered correctly is modeled as a function of item difficulty and person ability. The person ability, or ***person stage of performance***, stands for how well the person performs at the set of tasks. The items difficulty, or ***Rasch scaled item difficulty***, is how difficult items were empirically. Both stages of performance scales are based solely on whether or not a given order of hierarchical complexity is correctly carried out. The order of hierarchical complexity is the theoretical difficulty of the items. It has been found that item ***order of hierarchical complexity*** accounted for over 90% variance of the Rasch scaled item difficulty (Commons, Goodheart et al. 2008; Bernholt, 2009).

Argument for gaps between orders

The result from past empirical research showed that the orders of hierarchical complexity of items predicted the relative difficulty of the items in the survey (Commons, Goodheart, Pekker, Dawson, Draney, and Adams, 2008). This is evidence that order of hierarchical complexity is an ordinal scale.

In addition, Rasch variable map of item difficulty showed "gaps" between items at each order of hierarchical complexity—items at each order clustered together. Rasch scaled item difficulty of items "jumps" from the highest items of order n to the lowest item of order (n + 1), instead of increasing in a continuous motion.

Whether "gaps" really exist is of interest to the authors, because it may provide evidence that there are qualitative changes between different orders of hierarchical complexity. In the model of hierarchical complexity, tasks at each order of hierarchical complexity have a distinct order of difficulty. From each order of hierarchical complexity to the next, the demand of solving the task jumps in an ordinal manner. Therefore, items at the same order should be similar, or theoretically have the same difficulty. In real data, the items at each order always have a range due to noise. However, the authors expect that as sample size increase, Rasch scaled item difficulty of items at the same order will converge.

This paper will analyze the data from an instrument and analyze whether "gaps" exists between orders and if there is equal spacing between orders.

Arguments for Linearity and Equal Spacing

The second question is whether the distance of item difficulty between each order of hierarchical complexity is the same, or whether the scale is a linear scale. As far as the authors know, there are no equally spaced ordinal scales. As part of the theory of measurement in mathematical psychology, the model of hierarchical complexity might define this new kind of scale.

The order of hierarchical complexity is not an interval scale, because orders cannot be added (Luce and Tukey, 1964). If one adds a task at order 7 and a task at order 8, it does not produce a new task at order 15. However, it might be a linear, equally spaced scale.

If the order of hierarchical complexity is linear and equally spaced scale, then it must satisfy this relationship: $a * f(n) = f(a * n)$, where a is a constant, n is the order hierarchical complexity of an item, and f is a function that calculates the difficulty of the item. In addition, every jump of one order represents the same increase in difficulty.

Three Arguments to Support Equal Spacing

There will be three arguments given to support the likelihood of equal spacing: Fractal nature of the transition steps in performance; pattern recognition of two forms, stacked neural networks and same Diffusion process in recognition task at different stages.

A first form of support for equal spacings is found from the fractal nature of stage transition. ***Stage transition*** may be directly tied to the notion of order of hierarchical complexity. The next order behavior is defined in terms of two or more the next lower order behaviors, and the higher order task actions have to organize the lower order ones in a non-arbitrary way (Commons, Gane-McCalla, Barker and Li, in press). So stage transition is complete when the next order task actions have successfully addressed the next order tasks. What underlies stage transition is similar to what underlies stage itself in one major respect. In both cases, there are task that must be completed correctly. In stages of performance, the tasks successfully completed have a particular order of hierarchical complexity. In the performance version of transition, the performance is at a given ordered step along the way in transition. Both the order of hierarchical complexity and step orders are ordinals. The performance should reflect those orders. If going from one stage to the next, always follows the same process, then the stages are fractal in nature. This is because going from successfully addressing one order's tasks and then doing the next follows the same dynamical rules (Ross, 2008).

Just as the orders of hierarchical complexity are an ordinal scale, so also are the transition steps that individually comprise the transition sequence. Table 2 illustrates transition steps between stages. One way to visualize the relation of the transition step ordinal scale to the orders of hierarchical complexity is as follows. The orders of increasing hierarchical complexity are an ordinal scale (i.e., 0, 1, 2, 3, 4,…,15). The transition steps that lead from one order to another fall on another ordinal scale, which runs from 1 through 6. The ordinal nature means these are not like degrees of temperature that are on an equally spaced scale. Ordinal scales are simple counts of occurrences; in this case, task orders.

One major implication of this universal, self-similar pattern that shows up at all scales of tasks of any kind is, by definition, is that the stages described by the model of hierarchical complexity are fractal, as are the transition steps. That is, the same pattern repeats within each transition sequence and in more complex transition behaviors, fractals of the model's stage sequences also appear within transition sequences (Ross, 2008). "Formally, a fractal is an infinite, self-propagating pattern that repeats at every level of resolution" (Vicciardo, 2010). A mathematical fractal is based on an equation that undergoes iteration, a form of feedback based on recursion (Briggs, 1992). In the case of hierarchical complexity, a higher order action always results from the non-arbitrary organization of two or more, lower order actions. And, transition always follows the same pattern of rejection of previous action, followed by alternation of behaviors, followed by arbitrary behavioral combinations, and finally, by the non-arbitrary organization of behavior at the next higher order of complexity. Table 1 is an abbreviated description of the eight ordinally scaled transition steps. Because these steps occur between each stage, the spacing of the stages might be equal. Therefore, the assumption is that each coordination is of equal difficulty.

The way the ordinal scale of hierarchical complexity is generated in Table 1.

Let n = the order of hierarchical complexity

If an action at order $n + 2$ is one order higher than an action of order $n + 1$, and an action at order $n + 1$ is one order higher than an action at order n, then these actions of orders of hierarchical complexity can be generalized for all orders by using mathematical induction. Note that $n + 2$, $n + 1$, n is an ordinal sequence. If there were additional orders in between any two orders, $n + 1$ and n, they would have to meet the axioms, especially the three main ones, that higher order of complexity actions are defined in terms of the next lower order, organize them, and in a non-arbitrary way. Because that has been tested for over 50 sequences, and the commonalities of the characteristics of the task actions at those orders have been abstracted, we are very confident that there are no other orders except above 15.

A Second and Third Form of Support for Equal Spacing Come from Pattern Recognition

Stage change occurs when the actions of combining of the lower order entities into the higher order ones occurs successfully. When this is achieved, stage transition has been completed. We assert that all stage change requires pattern recognition. The pattern to be recognized is the ordering of lower order actions that works in the new stage. However, not all pattern recognition is stage change, but all pattern recognition occurs at some stages. Some examples that pattern

Table 2

Transition Steps Between Stages

Step	Step Name	Substep Name	Relations	Dialectical Form
1	*A*		Extinction of thesis from previous stage begins	Previous stage action does not solve many tasks. (Deconstruction begins) Extinction Process
2	*B*		Antithesis: Negation or complementation	Negation or complementation, Inversion, or alternate thesis
3	*A* or *B*		Relativism: Alternation of thesis and antithesis	Alternation of thesis and antithesis. There is no coordination of them
4	*A* and *B*		Smash: Synthesis begin	Unordered synthesis of components from *A* and *B*
		Smash 1	Random Hits, False alarms and Misses and Correct Rejections	Synthesis of components from *A* and *B* in a non-random order
		Smash 2	More Hits, lower Misses, excess False Alarms	Incorporates subsets producing hits at Stage ***n***. Basis for exclusion not sharp. (Overgeneralization.)
		Smash3	Correct Rejections increase , Excess misses, Lower Hits and False Alarms	Incorporates subsets producing correct rejections. Basis for inclusion not sharp. (Undergeneralization)
5	*A with B*		Synthesis and new thesis: New temporary equilibrium	Temporary equilibrium (synthesis and new thesis)

recognition is the same independent of stage are presented next. If the same mechanism is used over and over and the parameters in the models are roughly constant, it supports the notion that the stages will be equally spaced.

A second form of support for equal spacings is found in the neural network model for hierarchical complexity (Commons, 2008b). It is another model for pattern recognition. This model also shows that going from one stage to another follows the same process. The stages are modeled by stacking neural networks into layers. Neural networks at each stage are the same. Each network recognizes a pattern and stacks of neural network follow stages (Commons, 2008b). It can be established from the stacked neural network model that same steps or processes occur between each stage which indicates that the spacing between the stages might be equal.

A third form of equal spacing follows from evidence that the same process is used for stage change no matter what the stage. Further substantiation of the notion that transition between stages follows the same dynamic processes between every stage can be attained from Ratcliff, Thapar and McKoon's (2006) study. They applied a diffusion model in analyzing performance of participants of three different age groups on four different tasks. Ratcliff's (1978) diffusion model was first used as a theory of memory retrieval. He later used the model to analyze individuals' cognitive processes while making simple two-choice decision tasks. We suggest that the Ratcliff et al. (2006) study was about general pattern recognition. All four tasks used in the study required the participants to recognize a certain pattern before coming to a conclusion about the answer. The tasks were signal detection, letter discrimination brightness discrimination and memory recognition (see Appendix A for description of each task). Signal detection required recognition of numerosity, letter discrimination required recognition of letters, brightness discrimination required recognition of light intensity and memory recognition required recognition of words. It can be inferred that the reason Ratcliff et al. found that there were correlations across tasks in component processes for individual subjects was that this common underlying mechanism of pattern recognition is required in all the tasks.

At the same time, in coding these tasks in terms of hierarchical complexity, we have found that they differ. The tasks were scored and agreed upon by four trained scorers using hierarchical complexity Scoring System (HCSS) (Commons, Miller, Goodheart, and Danaher-Gilpin, 2005). The order of each task defers depending on whether the tests are administered on humans or animals. Signal detection requires participants to estimate as they are asked to decide whether the number of asterisks flashed on the screen is large or small. For animals, this is a minimum perceptual task which can be achieved at stage 2. However, since the participants were human in the study done by Ratcliff et al., the task was beyond a perceptual one as humans have the concept of what is large and small and they use the concept while making a decision about the number of asterisks. Those who successfully solve this task are performing at Stage 3. Similarly, letter discrimination is an order 3 task for animals such as pigeons. This task requires matching of letters. Letters are arbitrary symbols for animals which have no meaning. However, humans who successfully solve this task function at stage 6 because letters are not arbitrary symbols for them and processing letters requires one to function at a higher stage. Brightness discrimination is another perceptual task. Merely seeing the difference in brightness requires the participant to be at stage 1, choosing whether the stimulus is bright or dark requires one to function at stage 2, but saying whether the stimulus is dark or bright requires one to be at stage 4. Memory recognition is an order 8 task since participants have to discriminate words. Their familiarity with the words, in particular knowing about one word with respect to others, would affect their responses. It was not merely a matching task. Although these tasks differ in stages, they all require pattern recognition which works the same way at every stage. Only the number of layers changes at each stage. The highest stage always requires all the lower stage processing. The fact that the tasks differ in their orders of hierarchical complexity but require the same underlying process of pattern recognition suggests that, first of all, the mechanism of moving from one stage to the next is always engaging in pattern recognition. If so, this would suggest that such stage transitions are equally difficult and that the stages and the orders are equally spaced.

Current Paper

In the current paper, we explore and present evidence from empirical data to support two hypotheses. First, we hypothesize that there are "gaps" between items on different stages of performance.

We will test this hypothesis by analyzing the ***gaps***, which are difference scores of Lowest Rasch Item Difficulty score of the next higher order of hierarchical complexity items and the Highest Rasch scaled item difficulty score of items from the next lower order of hierarchical complexity items. The null hypothesis is that the average of size of gaps between items from contiguous orders of hierarchical complexity are equal to the average difference between Rasch scaled item difficulty of the more difficult item and that of the next less difficult item. We expect that the null hypothesis will be rejected.

Second, we hypothesize that there is equal spacing and linearity between the orders of hierarchical complexity. We test this hypothesis by running several different analyses such as simple linear regression, lack of fit test, *t*-tests and perturbations.

Method

Participants

There were 113 participants. Participants were recruited from online Listservs. All participation was voluntary and no compensation was given. Of these participants 47 (41.6%) were men and 66 (58.4%) were women. Self-reported ages ranged from 18 to 70 (M = 34.67, (SD = 13.76). Five participants' data of age had apparent mistakes and their data was not taken into account in the calculation of age demographics. Education varied from high school to graduate degree (35 high school graduates, 57 Bachelor's degree holders, 8 master's level degree holders and 13 doctoral level degree holders. M = Bachelor's degree).

Instruments

This study used the laundry instrument that was based on the Inhelder and Piaget's (1958) pendulum task. The laundry instrument asked participants whether or not a piece of laundry would be clean after varying treatment. Participants were required to view a table depicting what had already happened (informational episodes) and then make predictions about what would happen in a new episode. The instrument was in English and it was given in the United States. Each instrument that the participants received included tasks at the primary, concrete, abstract, formal, and systematic order in the model of hierarchical complexity.

Tasks. The history of the different variants arising from the pendulum task (Inhelder and Piaget, 1958) begins with the plant problem created by Kuhn and Brannock (1977) (also see Kuhn, 1974; Kuhn, and Angelev, 1976). Kuhn and Brannock used the plant problem because they felt it offered greater external and ecological validity than Inhelder and Piaget's pendulum task. To perform at the formal stage, the pendulum task required participants to perform an experiment by manipulating a single variable while holding all other variables constant. They had to figure out which variable controlled the rate that a pendulum weight would cross the low point. The content was is the physics domain, with which many participants were unfamiliar. The plant problem overcame this by offering various observations of what made a plant healthy or sick, which differed in several areas and required participants to make inferences based on numerous observations. Kuhn and Brannock (1977) felt that their plant problem more closely reflected "natural experiments" where the individual does not have to solve a controlled laboratory type experiment like the pendulum task.

Although the general direction of the changes was positive in terms of bringing greater ecological validity to studies of isolation of variables problems, the new instrument had certain issues such as the possibility for a participant to find multiple simple answers and a lack of consistency of the number of variables between episodes. Also, the participants were interviewed after which their replies were scored. As was the Piagetian (Inhelder and Piaget, 1958) tradition, single tasks were used to measure multiple stages for performance. Without a true independent variable, this made it difficult to see what contributed to the differences in performance. One had to make inferences from often incomplete interviews insufficiently probed. This step added a layer of confusion that would not exist if the outcome alone was scored. For this reason the original plant problem was altered by Commons, Miller, and Kuhn (1982). The new instrument had two positive and two negative possible causes in each episode. There were three episodes positive outcomes and three with negative ones. This created six information episodes that had four variables and ten test episodes. The combinations of three of the variables that were the compliment of the one that was obviously causal were also causal. However, for any finite number of trials, the combination of the complementary variables of the causal variable would also be causal. If the plant food were causal ingredient, then the combination of the leaf lotion, small or large pot, lot or a little water would also be causal. Almost no participants detected this combination.

For the current study, the tasks' properties that were either truly independent variables or quasi-independent variables were separated from performance on those tasks. This allowed us to see that the difficulty of the items as represented by their order of hierarchical complexity was the most predictive task variable. The results from these studies support the model of hierarchical complexity's effectiveness in predicting stage of performance on tasks of differing hierarchical complexity.

There was one additional problem with these earlier versions of the isolation of variables problems. Only one version of the problem was ever administered. It had different number of variables per episode and one outcome per episode that was supposed to be correct. Explanations of choices were scored as to stage from the inferred task that the participants were successfully addressing. Different participants were taking into account different amounts of information given at different stages (Miller, personal communication, June 20, 2010). Some appeared to not attend to or refer too much, if any, of the given information. Others focused on only one or two variables, not checking each variable systematically. Of course, some properly attended to the variables eliminating the non-causal ones (Formal Stage 10). A few saw more complex relationships, including the possible interactions between instances in their responses (Concrete Stage 8). Clearly, those who did not refer too much of the information, or only considered one variable were scored lower than those who considered all four. In placing participants into the different "stage" categories there was always uncertainty as to whether that participant really belonged there or whether their response might be due to other factors that were not controlled. As investigators, we began to consider what might happen if we created different versions of the isolation of variables problem, some of which were simpler and some of which were more complex.

Instrument description. This section describes the format of the tasks. A task is what the participant was required to complete in order to answer a problem correctly. Participants presented a simple deduction, A leads to B, and A is presented, they were asked to predict whether B or not B was correct. Although the content, country, and language of each instrument varied, the tasks requiring successful completion in order to answer the problem correctly were virtually the same across instruments. The following descriptions of the tasks for each order are taken from the Decision Making Instrument. From the Rasch analysis of the combined data, the item reliability was 1.00. The following descriptions of the tasks for each order were taken from the laundry instrument.

Primary. At the primary order 7, there were two informational episodes each showing one ingredient and an outcome. One outcome showed a case where the stain was removed and the other showed that it was not. There was just a single ingredient variable in both cases. The prediction episode repeated the one and only ingredient in the informational episode and asked the participant to predict the outcome, which was described verbatim in the informational episode. The participant was tasked with recognizing that the answer was directly given in the informational episodes by each ingredient – outcome pair. In all the orders, the predictions were based on bidirectional associations of ingredients and outcomes so that not only was "if A then B" was true but also "A if and only if B" was true.

Concrete. At the concrete order 8, two or more instances of the primary order actions were combined. At the primary order, a single ingredient was the predictor of the correct outcome. At the concrete order, there were multiple ingredients as possible predictors, but only one that was a correct predictor of the correct outcome. The prediction episode repeated the multiple ingredients from the informational episode and the participant had to choose that the stain would behave the way it did in the informational episode. This required them to consider not just one ingredient – outcome pair but two or more ingredient - outcome pairs. The participants were given the task of recognizing that the answer was directly given in the informational episodes. Two versions of the concrete order 8 tasks were used. The laundry instrument had four informational episodes that showed whether or not *four* ingredients removed a stain. Subsequent versions of the instruments use four informational episodes that show whether or not *two* ingredients together produce a clean or dirty cloth.

*Abstract***.** At the abstract order 9, there were three informational episodes that showed whether or not *two* ingredients removed a stain. The abstract order was the first order that required participants to isolate a variable in order to determine the outcome of the prediction episode. But there were only two variables given, so the task required a minimal amount of isolation of variables. The participant was required to examine one variable at a time in order to find a consistent outcome. For example, in a particular problem, type of bleach determined the effectiveness of the combination. In another problem, water $_{\text{temperature}}$ was predictive. The process combined two concrete order tasks, by first having to find which variable did not predict the outcome and then finding which variable predicted the outcome. Looking at one variable at a time was the organizing action.

Formal. At the formal order 10, there were six informational episodes with four variables within each episode that showed whether or not one of the *four* ingredients removed a stain. In the prediction episodes, participants had to isolate a single ingredient (variable) whose value predicted whether that particular ingredient would clean clothing irrespective of the values of the other ingredients. Many of the combinations presented in the prediction episodes did match combinations presented in the informational episodes. The abstract order of isolation of variables action therefore had to be applied in a systematic way to each of the variable outcome instances across episodes. It is the antecedent-consequent testing of the variables that orders the abstract order 9 of testing of single variables. Therefore proficiency a*t*-testing one variable at a time and identifying univariate predictive relations must be acquired in order to solve the more complex problems.

*Systematic***.** At the systematic order 11, the informational episodes stated, which are *combinations* of two of four types of ingredients produced clean or dirty clothing. In the prediction episodes, participants identified the two variables whose values predicted whether or not particular combinations of ingredients would clean clothing. For example, in a particular problem, type of bleach ***and*** water temperature determined the effectiveness of the combination. In another problem, type of detergent OR type of booster were predictive. This conception of causality was multivariate: If (A and B), then C, or if (A or B), then C. Either type of rule could govern causal relations at the systematic order. Systematic-order rules organize and coordinate formal-order rules. The systematic order 11 action of testing combinations of variables with different combination rules orders the formal order 10 action of systematically testing a single variable.

Procedure

The instrument was presented in a survey form online. The instrument was distributed among various internet groups and Listservs. The tasks were presented in a sequence from easy to hard. Having done the easy problems, which most of the participants did, served as a support for the harder performance. There is a large literature that shows that one gets better at performance in going from easy to hard (Aamodt and Mcshane, 1992; Hodson, 2006). The items were coded as correct or incorrect with missing answers being assumed incorrect.

Data was analyzed using the Rasch model. The Rasch analysis yields two scales, the person stage of performance and the Rasch scaled item difficulty. The Rasch scaled item difficulty will be analyzed in this paper. First, independent sample *t*-tests were used to test the existence of gaps between two adjacent orders. It tested whether these gaps are statistically significant. Secondly, several analyses were done to test whether the order of hierarchical complexity is a linear and equally spaced scale. Analysis that were used are simple linear regression model, lack of fit test, a test on whether spacings are equally spaced and a test on how much we can perturb the linearity of the order of hierarchical complexity before significantly reducing its predicting power.

Results

Rasch Analysis

A Rasch analysis (Linacre, 2009, Adams and Khoo, 1993) was performed. Rasch analysis is a method for obtaining objective, fundamental, linear measures (qualified by standard errors and quality-control fit statistics) from stochastic observations of ordered category responses (Wright and Stone, 1979). It used logistic regression that serves to minimize the errors in person and item scores. Rasch analysis then takes the raw person and item scores and converts them into equal interval linear scales. The item scores, the Rasch scaled item difficulty, represent how difficult the item was. The person scores represent how well a person dealt with the item difficulty. The Rasch scaled item difficulty was used to estimate linear measures. With Rasch analysis, these measures are item-free (item-distribution-free) and person-free (person-distribution-free). This means that the measures are statistically equivalent for the items regardless of which persons (from the same population) are analyzed, and for the people regardless of which items (from the same set) are analyzed. Analysis of the data at the response-level indicates to what extent these ideals are realized within any particular data set. The higher a person's performance score relative to the difficulty of an item, the higher the probability of a correct response on that item by the participant. When a person's location on the latent trait is equal to the difficulty of the item, by definition, there is a 0.5 probability of a correct response.

This paper only analyzed the Rasch scaled item difficulty. The result of Rasch analysis shows that the range of the Rasch scaled item difficulty was from −4.56 (Primary 7) to 3.94 (Systematic 11). The higher an item is on the scale, the more difficult the task. See Figure 2 of the Rasch Variable Map. On the right side of the scale are items. The letters stand for the orders of hierarchical complexity of the items (P—Primary 7, C—Concrete 8, A—Abstract 9, F—Formal 10, S - systematic 11). The Rasch variable map shows that mean Rasch scaled item difficulty for a given order of hierarchical complexity is sequenced in the same way the orders of hierarchical complexity are as shown in Figure 2, the Rasch map. In addition, there were no items that were out of order as well. These conditions satisfy the weak and intermediate condition for an ordinal scale.

Test for Gaps

By visual inspection of the Rasch variable map (see Figure 2), there are gaps between two adjacent orders of hierarchical complexity. To answer the questions more precisely about whether or not there are gaps between items of adjacent orders of hierarchical complexity, an independent sample *t*-test was used. It tested whether the gaps were significant. Some definitions are set forth.

Gap = Lowest Rasch scaled item difficulty score of the next higher order of hierarchical complexity items minus the Highest Rasch scaled item difficulty score of items from the next lower order of hierarchical complexity items.

Item break = Rasch scaled item difficulty of the more difficult item—Rasch scaled item difficulty of the next less difficult item.

The null hypothesis was that the average size of gaps between items from adjacent orders of hierarchical complexity are equal to the average difference in Rasch item difficulty for the lower of the adjacent hierarchical complexity groups. A model was constructed to test the hypothesis. For details of how the model was constructed, please see Appendix B.

Let i = the observation number, which goes from 1 to 102.

$DR_i = \beta + a_7\, I7_i + a_8\, I8_i + a_9\, I9_i + a_{10}\, I10_i + a_{11}\, I11_i + \varepsilon_i$ where

DR_i = the difference of Rasch scaled item difficulty between item I and item $(i-1)$

```
     Participants MAP OF RANKS
              <more>|<more difficult>
   7                +
                    |
                    |
               # T  |
   6                +
                    |
            ##      |
                    |
   5          .##   +
               .##  |T
                .#  |
               ###  |
   4        .######  +  sy
              ####  |
                  S|  sy sy
               ##   |  sy sy sy
   3           ###  +
                ##  |  sy sy sy  sy sy
                #   |  sy sy sy  sy sy sy sy (Systemtatic)
               .##  |S sy sy sy  sy sy sy sy sy
   2          ###   +  sy sy sy  sy
         ########   |
              ###   |  f1 f1 f1 f1
              ###   |  f1 f1 f1 f1  f2  f2  f2  f2  f2  f2 f2  f2 (formal)
   1           ##   +  f1 f1 f2 f2
                #   |
                  M|
         ########   |
   0        .#####  +M ab ab
               ##   |  ab ab ab ab  ab  ab  ab  ab  ab (abstract)
                #   |  ab ab ab ab  ab  ab
              .##   |  ab ab ab
  -1            .   +
                    |
         .#######   |
               .#   |  cr cr cr cr  cr
  -2            .   +  cr cr                       (concrete)
                . S|S cr cr cr cr
                .   |  cr
       .#########   |
  -3        .####   +  pr
              .##   |  pr pr pr pr
              .##   |  pr pr pr pr pr pr pr pr pr pr (primary)
                #   |  pr pr
  -4            #   +  pr
                #   |
                    |  pr pr
                .   |T
  -5            . T+
                .   |
                    |
                    |
  -6            .   +
              <less>|<easier>
```

Figure 2. Rasch Variable Map of laundry Instrument

β = the average of gaps

a_n = the difference between the average of item break at order n and the average of Gap β

In_i = {1,0} {is, is not} a difference in Rasch scores for Hierarchical order or group *n*

ε_i is a random variable fulfilling the Gauss Markov conditions.

The results showed that indeed, there was a significant sized Gap. After data was fit to the model:

$$DR = 0.65500 - 0.57447\ I7i - 0.58864\ I8i - 0.60553\ I9i - 0.62237\ I10i - 0.58397\ I11i$$

This equation shows that the average of gaps was 0.655. The average item break at each stage was smaller than the average gap size as shown by the a_n being negative.

The null hypothesis was that the average gaps between contiguous orders of hierarchical complexity are equal to the item breaks within each order. There were 5 null hypotheses: $a_n = 0$, n = 7, 8, 9, 10, and 11. The alternative hypothesis was that the average size of gap was bigger or smaller than the average difference in Rash item difficulty scores. There were 5 alternative hypotheses: $a_n \neq 0$, n = 7, 8, 9, 10, 11.

Five t_n-tests were used to test the nulls against the alternative. For the formula for the *t*-test, please refer to the Appendix B. The result of the tests showed:

t_7 (97) = −10.014, $p < 2^{-16} \approx 0.00000$

t_8 (97) = −9.667, $p < 2^{-16} \approx 0.00000$

t_9 (97) = −10.555, $p < 2^{-16} \approx 0.00000$

t_{10} (97) = −10.848, $p < 2^{-16} \approx 0.00000$

t^{11} (97) = −10.499, $p < 2^{-16} \approx 0.00000$.

The null hypothesis was rejected in all the tests implying that the items breaks were significantly different from the gaps. In addition, because the average of Items Breaks at each order was smaller than the average of gaps, average item breaks were significantly smaller than the average gaps. Therefore, this shows that gaps exist.

Test for Linearity and Equal Spacing

This section investigates whether the order of hierarchical complexity was a linear and equally spaced scale. If the Rasch scaled item difficulty of task items has equal spacing, it implies that the orders of hierarchical complexity had to have equal spacing. What else could have produced the equal spacing? We tested a number of ways to look at the possibility of non-linearity. We recognize that it is impossible to prove linearity by failing to reject the null hypothesis. In all cases, we failed to reject the null hypothesis that things were linear. As far as we know, if one fails to reject null hypothesis, one can still hold the assumption of linearity. The following results failed to provide evidence to show that the scale is not linear. There were four tests: 1) a simple regression model was constructed; 2) a lack of fit test shows that the linear regression model explains as much variance as the separate means model, indicating that linearity cannot be rejected; 3) a test on the spacing between Rasch scaled item difficulty are done, showing that equal spacing cannot be rejected; 4) we perturbed the linear order of hierarchical complexity. This was another way of testing whether the model of hierarchical complexity was linear. This was to test how well the linear regression predicts Rasch scaled item difficulty when the Task order of hierarchical complexity was perturbed, or noise was added to the scale.

Simple Linear Regression. First, a linear regression model was constructed. The dependent variable was the Rasch scaled item difficulty. The independent variable was the item order of hierarchical complexity. The parameters were the slope of the straight line function, b and the intercept, a.

Rasch scaled item difficulty = $\alpha + \beta * n$; n = Item order of hierarchical complexity, α = intercept, b was the slope, and n = 7, 8, 9, 10, 11. The linear regression model shows that task order of hierarchical complexity significantly predicted Rasch scaled item difficulty, r (98) = .983, r^2 = .975, $p < 0.001$. The size of the variance explained by the item order of hierarchical complexity, an ordinal scale, showed that the linear scale was highly predictive of the Rasch scaled item difficulty. See Figure 3 for the regression line.

To securitize the result, the residual graph was shown in Figure 4. It shows that for task orders 7, 8, 9, 10 and 11, the mean of the residuals center around zero, and there is no indication of systematic relationship between the residuals and the task order of hierarchical complexity.

Test for Linearity. A lack of fit test was used to analyze whether the relationship between orders of hierarchical complexity and Rasch scaled item difficulty was linear. The lack of fit test compares the residuals of the linear regression model to the separate means model. When the linear regression model explains significantly less variation of the dependent variable than the separate means model, it is usually an indicator that the linear regression model is not a good fit to the data. The relationship between the independent variable and dependent variable is not linear (Ramsey and Schafer, 2012).

The null hypothesis was the linear regression model explains significantly less variance than the separate means model. The alternative hypothesis was the linear regression model and the separate means model explains equal amount of variance in the data.

For detailed procedures of the test, please see Appendix C. The lack of fit test shows that $F(3) = 1.944$, $p = 0.128$. The separate means model did not explain significantly more variance than the linear regression model. The null hypothesis was that the spacing was unequal was not rejected. The result indicated that the linear relationship between the Task order of hierarchical complexity and the Rasch scaled item difficulty was not rejected by this analysis. The linearity assumption still held.

Test for Equal Spacing. Using a *t*-test, this analysis tests whether there are equal spacing between adjacent orders of hierarchical complexity. ***Spacing*** is defined as the increment from the average of Rasch Scaled Item Difficulties of a lower order to the average of Rasch scaled item difficulty of the next higher order. There are four spacings as there are five orders of hierarchical

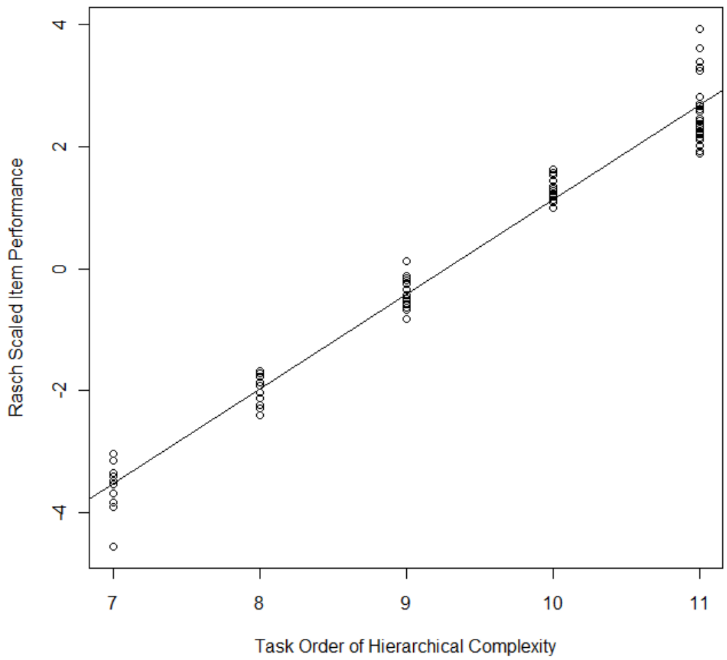

Figure 3. Simple Linear Regression of laundry Data

complexity. The result of this analysis showed that we cannot reject the null hypothesis of equal spacing. The following is a brief explanation of the analysis. For complete steps of the calculation, please refer to Appendix D.

The following linear regression model was constructed:

R_D = Rasch scaled item difficulty = $\beta_7 + \gamma_8 I_{8i} + \gamma_9 I_{9i} + \gamma_{10} I_{10i} + \gamma_{11} I_{11i} + \varepsilon_i$

R_D = Rasch scaled item difficulty;

I_{ni} = {1, 0} when the item {is, is not} at the order of hierarchical complexity denoted by *n*. n = {7, 8, 9, 10, 11};

β_7 = is the average value of the Rasch scaled item difficulty for items in order 7.

γ_8 = the estimate of the difference between the average Rasch scaled item difficulty at order 8 score and average Rasch scaled item difficulty at order 7 score

Accordingly, ($\beta_7 + \gamma_8$) estimated the average Rasch scaled item difficulty at order 8. Similarly, {γ_9, γ_{10}, γ_{11}} estimated the difference between the average Rasch scaled item difficulty at order {9, 10, 11} score and the Rasch scaled item difficulty at order 7. Accordingly, {$\beta_7 + \gamma_9$, $\beta_7 + \gamma_{10}$, $\beta_7 + \gamma_{11}$} estimated the average Rasch scaled item difficulty for items at orders {9, 10, 11}.

The null hypothesis was that all spacings are the same. The alternative hypothesis was that at least one pair of spacings are different. For the purpose of calculation, the null hypothesis was broken into three parts.

H_{01}: The spacing between order 9 and 8 will be the same as the spacing between order 8 and 7.

Or $\gamma_9 - 2\gamma_8 = 0$.

H_{02}: The spacing between order 10 and 9 will be the same as the spacing between order 9 and 8.

Or, $\gamma_{10} - 2\gamma_9 + \gamma_8 = 0$

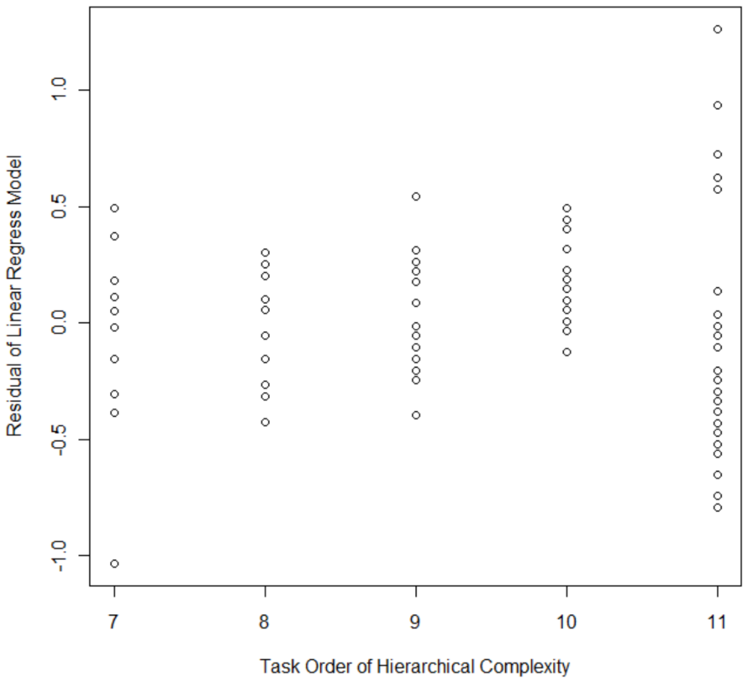

Figure 4. Residuals of the Simple Linear Regression of laundry Data

H_{03}: The spacing between order 11 and 10 will be the same as the spacing between order 10 and 9.

Or, $\gamma_{11} - 2\gamma_{10} + \gamma_9 = 0$

One sample *t*-tests were used to test these hypotheses. Please refer to Appendix D for the details of the computation. The result showed that we cannot reject any of these null hypotheses: 1) For H_{01}, t (97) = 0.240, p = 0.595; 2); For H_{02}, t (97) = 0.0526, p = 0.479; 3); lastly, for H_{03}, t (97) = 0.7949, p = 0.214. Therefore, we cannot reject the null hypotheses that all the spacing between the orders is the same. This result is consistent with the result of lack of fit test, which cannot reject linearity of the orders of hierarchical complexity.

Perturbations

As it is shown in previous analysis, linearity of order of hierarchical complexity cannot be rejected. This, however, does not strictly prove equal spacing. This section of the paper explores how much random noise needs to be added to the orders of hierarchical complexity in order to reject the linearity hypothesis. This gives an upper limit to how far from equal spacing and linearity the orders might be.

A computer randomization program generated a list of 0s and 1s. When 0 came up, 0.05 was subtracted from n^{th} order of hierarchical complexity. When 1 came up, 0.05 is added to the order *n*. For example, if 1 comes up, order 7 becomes order 6.95. If 0 comes up, order 7 becomes order 7.05. This procedure was applied to every order, but only once per order. The result orders from the perturbations were 6.95, 8.05, 9.05, 9.95 and 10.95.

Next, a linear regression of the Rasch scaled item difficulty was run on the newly defined order scale. What was of interest was the predictability of the new scale, or the value of *r*. The result showed that $r = 0.987$. The same procedure was repeated three more times. The *r*'s at the noise level of 0.05 are 0.987, 0.988, 0.988, 0.975, which gives an average of 0.9875 as seen in Figure 5. The reason to repeat the procedure and obtain the average is because what was of interest was the average predictability, or value of *r,* of the perturbed order of hierarchical complexity. Every time the procedure was applied, the resulting new scale is different. Multiple samples were obtained making the results closer to the real mean.

Next, the same procedure was applied using noise level 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45 and 0.5. The respective average of *r*'s of the linear regression models are 0.9875, 0.9865, 0.98425, 0.98075, 0.98225, 0.974, 0.9695, 0.96575, 0.97025, 0.94425. A quadratic linear regression was run on the *r* scores against the size of the perturbation. The result showed that the size of the perturbation significantly predicted the *r* scores with $r(9) = 0.933$, $r^2 = 0.871$.

In addition, it is found that perturbing the order of hierarchical complexity by more than 0.25 produced a significant difference in the predictability of the scale. This is more than 1/4 of an order. Using the Fisher *r*-to-*z* transformation, the significance of the difference between the *r* found in the original linear regression model and the *r*'s found in the new models when the order

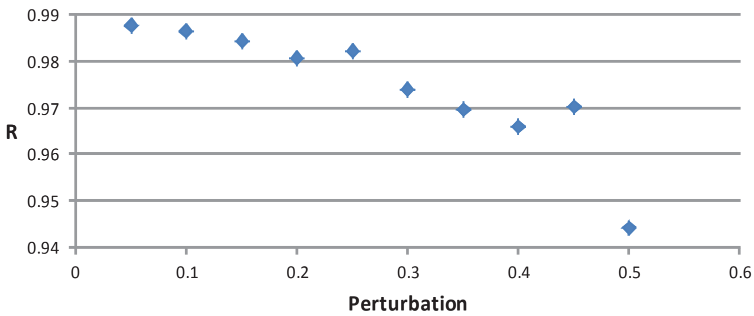

Figure 5. Size for perturbation to the order to hierarchical complexity versus R

of hierarchical complexity is perturbed, were assessed. When the noise was 0.25, the difference was significant at the 0.1 level, with $z = 1.68$, $p = 0.093$. When noise = 0.35, the difference was significant at 0.05 level, with $z = 2.74$, $p = 0.006$.

Discussion

The results indicated that there are gaps between stages as the null hypothesis was rejected in all the *t*-tests performed to test for existence of gaps. According to the results, the items breaks are significantly different from the gaps. In addition, because the average of Items Breaks at each order was smaller than the average of gaps, average item breaks were significantly smaller than the average gaps. Therefore, gaps and item breaks are not the same and gaps do exist. The existence of gaps shows that the ordinal nature of the scale is not just an assumption. The "stage process" reflects the ordinality of underlying tasks.

Results also supported the hypothesis that there is equal spacing between orders of Hierarchical Complexities and it is a linear scale. The simple linear regression analysis showed that the size of the variance explained by the item order of hierarchical complexity, an ordinal scale, shows that the linear scale was highly predictive of the Rasch scaled item difficulty. The lack of fit test showed that the linear regression model and the separate means model explained equal amount of variance in the data which indicates that the order of hierarchical complexity is a linear scale. Consistent with this result is also the result from the *t*-tests for equal spacing. We were unable to reject the null hypothesis that the spacing between the orders is the same. More evidence to substantiate the claim that there is equal spacing between orders comes from the results of the perturbation of linear order of hierarchical complexity. The results showed that the linear regression did predict Rasch scaled item difficulty when the Task order of hierarchical complexity was perturbed, or noise was added to the scale. Again, this supports the claim that there is equal spacing between orders of hierarchical complexity and that the scale is linear.

Implications

The fact that there are gaps between orders of hierarchical complexity and that the scale is equally spaced and linear implies that the difficulty of going to the next stage is the same regardless of what stage someone is performing at. This has large numbers of implications. This allows one to treat orders as actual numbers, and not just indication of relative position.

It might mean the order of hierarchical complexity, n, is a measure of the quantity of hierarchical information. Given that tasks at order $n + 1$ are defined by and coordinate two or more tasks at order n, the minimum number of order 1 tasks that an order n task is 2^n. Equal spacing might indicate that 2^n is well defined and therefore $\log 2^n = n$, a parallel notion to bits. That might mean that n is a measure of the quantity of hierarchical information and could be called Hbits.

If n is a measure of hierarchical information, it might be applied to reductionism in general. For example, the hierarchy of: 1) Strings which are combined to form; 2) Quarks; which are combined to form 3) Subatomic particles which are combined to form; 4) Atoms which are combined to form; 5) Molecules; 6) etc. Hierarchical information is gained as one moves up the hierarchy. The actions are the combining of the lower order entities into the higher order ones.

Future Research

Although gaps exist between adjacent order tasks, it does not mean that development does not take place within these gaps. Development does not show up as correct next order task actions. In addition to order task actions, there are subtask actions between orders (Commons, in press). *Subtask actions* organize only one action from the same order of hierarchical complexity and one or more from previous orders. They are necessary prerequisites to other same order tasks. *Subsubtask actions* coordinate actions from different orders that are precursors that are necessary for acquisition of subtask actions but are not necessary for performance after acquisition. Now that it has been established that there are gaps and equal spacing between orders of Hierarchical Complexities, a future direction for research would be to investigate how subtask and subsubtask actions work between stages.

References

Aamodt, M. G., and McShane, T. (1992). A meta-analytic investigation of the effect of various test item characteristics on test scores and test completion times. *Public Personnel Management*, *21*(2), 151-160.

Adams R. J., and Khoo, S. T. (1993). *Quest: The interactive test analysis system.* Hawthorn, Victoria: ACER.

Bernholt, S., Parchmann, L., and Commons, M. L. (2009). Modelling scientific competence between research and teaching practice. *Zeitschrift fur Didaktik der Naturwissenschaften*, *15*, 217-243.

Binder, C. (2000). *Component/composite analysis and programming from the "real world."* Paper presented at the Association for Behavior Analysis, Washington, DC.

Bradley R. A., and Terry M. E. (1952) Rank analysis of incomplete block designs, I. The method of pair comparisons. *Biometrika*, *39*. 324-345

Briggs, J. (1992). *Fractals: The patterns of chaos*. London, UK: Thames and Hudson.

Commons, M. L. (2001). A short history of the society for quantitative analyses of behavior. *The Behavioral Analyst Today. 2*(3). 275-279.

Commons, M. L. (2008a, July). The small effects of non-hierarchical complexity variables on performance. In Z. Pizlo (Chair). Symposium conducted at the meeting of Mathematical Psychology, Washington D.C.

Commons, M. L. (2008b). Stacked neural network must emulate evolution's hierarchical complexity. *World Futures, 64,* 444-451. doi:10.1080/02604020802301568

Commons, M. L. (in press). What are the relationships between four notions of stage change. *Journal of Adult Development.*

Commons, M. L., and Calnek, A. D. (1984). *The Genetic Epistemologist, 9*(2), 11-16.

Commons, M. L., Gane-McCalla, R., Barker, C. D., and Li, E. Y. (in press). The model of hierarchical complexity as a measurement system. *Journal of Adult Development*.

Commons, M. L., Goodheart, E. A., Pekker, A., Dawson, T. L., Draney, K., and Adams, K. M. (2008). Understanding Rasch measurement: Using Rasch scaled stage scores to validate orders of hierarchical complexity of balance beam task sequences. *Journal of Applied Measurement, 9*, 182-199.

Commons, M. L., Miller, P. M., Goodheart, E. A., and Danaher-Gilpin, D. (2005). Hierarchical complexity Scoring System: How to score anything. Retrieved from http://www.dareassociation.org/papers.php

Commons, M. L., Miller, P. M., and Kuhn, D. (1982). The relation between formal operational reasoning and academic course selection and performance among college freshmen and sophomores. *Journal of Applied Developmental Psychology, 3,* 1-10

Commons, M. L., and Pekker, A. (2008). Presenting the formal theory of hierarchical complexity. *World Futures: Journal of General Evolution, 65*(1-3), 375-382.

Commons, M. L., Rodriguez, J. A., Adams, K. M., Goodheart, E. A., Gutheil, T. G., and Cyr, E. D. (2006). Informed Consent: Do you know it when you see it? Evaluating the adequacy of patient consent and the value of a lawsuit. *Psychiatric Annals*, *36*, 430-435.

Dawson, T. L. (2002). A comparison of three developmental stage scoring systems. *Journal of Applied Measurement*, *3*, 146-189.

Dawson, T. L. (2003). A stage is a stage is a stage: A direct comparison of two scoring systems. *Journal of Genetic Psychology*, *164*, 335-364.

Hodson, D. (1984/2006). The effect of changes in item sequence on student performance in a multiple-choice chemistry test. *Journal of Research in Science Teaching. 5*(21), 489-495.

Inhelder, B., and Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* New York. NY: Basic Books, Inc.

Inhelder, B., and Piaget, J. (1969). *The early growth of logic in the child*. New York, NY: W. W. Norton.

Krantz, D. H., Luce, R. D, Suppes, P., and Tversky, A. (1971). *Foundations of measurement, Vol. I: Additive and polynomial representations.* New York, NY: Academic Press.

Krippendor, K. (2009). Mathematical theory of communication. In S. W. Littlejohn and K. A. Foss (Eds.). *Encyclopedia of communication theory* (pp. 614-618). Los Angeles, CA: Sage.

Kuhn, D. (1974). Inducing development experimentally: Comments on a research paradigm. *Developmental Psychology, 10* (5), 590-600.

Kuhn, D., and Angelev, J. (1976). An experimental study of the development of formal operational thought. *Child Development, 47,* 697-706.

Kuhn, D., and Brannock, J. (1977). Development of the isolation of variables scheme in experimental and "natural experiment" contexts. *Developmental Psychology, 13*(1), 9-14.

Kuhn, D., and Ho, V. (1977). The development of schemes for recognizing additive and alternative effects in a "natural experiment" context. *Developmental Psychology, 13*(5), 515-516.

Kuhn, D., Ho, V., and Adams, C. (1979). Formal reasoning among pre- and late-adolescents. *Child Development, 50,* 1128-1135.

Linacre, J. M. (2009). FACETS Rasch Software [Computer software]. Chicago, IL: Winsteps. Retrieved from http://www.winsteps. com/winpass.htm

Luce, R. D. (1959). *Individual choice behavior*. New York, NY: Wiley.

Luce, R. D., and Tukey, J. W. (1964). Simultaneous conjoint measurement: a new scale type of fundamental measurement. *Journal of Mathematical Psychology*, *1*, 1-27.

Ramsey, F. L., and Schafer, D. W. (2012).The statistical sleuth: A course in methods of data analysis. Boston, MA: Brooks/Cole.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago, IL: University of Chicago Press.)

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59-108.

Ratcliff, R., Thapar, A., and McKoon, G. (2004). A diffusion model analysis of the effects of aging on recognition memory. *Journal of Memory and Language*, *50*, 408-424.

Ratcliff, R., Thapar, A., and McKoon, G. (2006). Aging and individual differences in rapid two-choice decisions. *Psychonomic Bulletin and Review*, *13*, 626-635.

Ross, S. N. (2008). Fractal transition steps to fractal stages: The dynamics of evolution, II. *World Futures*, 64, 361-374.

Shannon, C. E., and Weaver, W. (1949). *The mathematical theory of communication*. Urbana, IL: University of Illinois Press.

Skoe, E. A. (In press). Measuring care-based moral development. *Journal of Adult Development.*

Vicciardo, M. (2010). *Looking at fractals, the universe*, ISBN 0-7167-1186-9

Wright, B. D., and Stone M. (1979). *Best test design*. Chicago, IL: MESA.

Appendix A

Two-decision tasks used by Ratcliff and McKoon (2006).

For each experiment, on alternating blocks of trials, instructions stressed that responses be either as accurate as possible or as fast as possible. The subjects were given feedback appropriate to the instructions, either accuracy feedback on each trial or a "too slow" message when an RT (Response Time) was over 700 msec. Responses were given by using specific keys on the keyboard.

Signal Detection

For each trial, a number of asterisks between 1 and 100 was generated from a signal distribution, normal with mean 57.5, or a noise distribution, normal with mean 39.5, each with an SD of 14.4. The asterisks were placed in random positions in a 10 X 10 array of blank characters on a computer screen. The subjects were asked to decide whether the number of displayed asterisks was "large" or "small." Accuracy feedback was given on all trials: If the number of asterisks was very large or very small, feedback indicated that "large" or "small," respectively, was the correct response. For intermediate numbers of asterisks, feedback was probabilistic, sometimes indicating "large" and sometimes "small" as the correct response. There were 12 blocks of 96 trials per session. For the data analyses, the numbers of asterisks were grouped into eight experimental conditions so that the mean RTs and accuracy values were about the same for the stimuli within a group.

Letter Discrimination

For each block of trials, there were two target letters continuously displayed in the top left and right corners of the computer screen. On each trial, one of the letters was displayed at the center of the screen for 10, 20, 30, 40, 50, or 60 msec and then masked. A subject's task was to indicate which letter was presented. There were 12 blocks of 96 trials per session. Performance for the 40-, 50-, and 60-msec durations was near ceiling, so data from these conditions were combined into one condition for data analyses.

Brightness Discrimination

The stimuli were 64 X 64 squares of black and white pixels displayed on a gray background of 320 X 200 pixels. There were six levels of brightness for the squares, achieved with six values of the probability of a pixel being white (.350, .425, .475, .525, .575, and .650). A square was displayed for 50, 100, or 150 msec, followed by a mask made up of four 64 X 64 checkerboard patterns presented sequentially, and the subjects were asked to decide whether each square was "bright" or "dark." There were eight blocks of 144 trials per session.

Recognition memory

The stimuli were high-, low-, and very-low-frequency words (Ratcliff, Thapar, and McKoon, 2004) in 20 study-test blocks per session. For each block, the study list consisted of words displayed for 1 sec each, 9 presented once and 9 presented three times (3 high, 3 low, and 3 very low frequency in each case), and the immediately following test list consisted of the 18 studied words plus 18 new words (6 high, 6 low, and 6 very low frequency). For each session, stimuli were chosen randomly without replacement from the three pools.

Appendix B
Test for gaps

The following model is proposed to answer the question, "Is there really a gap?"

$$DR_i = \beta\, J_i + \gamma 7\ I\ 7_i + \gamma 8\ I8_i + \gamma 9\ I9v + \gamma 10\ I10_i + \gamma 11\ I11_i + \varepsilon_i \qquad [1]$$

Let i = the observation number, it goes from 1 to 100. It used to define the difference between Rasch item scores within a group

DR_i = Rash Scaled Item Difficulty i - Rasch scaled item difficulty $i - 1$. The Rasch scores are sorted from smallest at the bottom to largest at the top

J_i = {1,0} according as the observation I {is, is not} a gaps difference;

I indicates which group a difference in item score belongs to (that is, group 7, group 8, group 9, group 10 or group 11); that is, it represents the difference score within each difficulty group (7, 8, and so on).

n is the order [or group, that is, 7, 8, 9, 10, 11]

In_i = {1,0} {is, is not} a difference in Rasch scores for Hierarchical order or group n;

β is the Average of the gaps scores;

γ (gamma) or rather γ_n is the Average continuous difference score in Hierarchical level n;

Epsilon$_i$, ε_i , is a random variable fulfilling the Gauss Markov conditions

There is a linear dependency in this model, namely $J_i + I7_i + I8_i + I9_i + I10_i + I11_i = 1$

To eliminate this linear dependency, solve for J_i: $J_i = 1 - (I7_i + I8_i + I9_i + I10_i + I11_i)$

Substitute the value for J_i into the model (equation [1]) and combine like terms. Equation [2] results:

$$DR_i = \beta + (\gamma 7 - \beta)\, I7_i + (\gamma 8 - \beta)\, I8_i + (\gamma 9 - \beta)\, I9_i + (\gamma 10 - \beta)\, I10_i + (\gamma 11 - \beta) I11_i + \varepsilon_i \qquad [2]$$

Let $a_n = (\gamma n - \beta)$, and let a_n denote the least squares estimate of a_n

$$DR_i = \beta + a_7\, I7_i + a_8\, I8_i + a_9\, I9_i + a_{10}\, I10_i + a_{11}\, I11_i + \varepsilon_i \qquad [3]$$

After data is fit to the model,

$$DR = 0.65500 - 0.57447\ \mathrm{I7i} - 0.58864\ I8_i - 0.60553\ I9_i - 0.62237\ I10_i - 0.58397\ I11_i$$

This equation shows that the average of gaps is 0.655. All a_n's are negative, which shows that the average item break at each stage is smaller than the average gaps.

The null hypothesis is that the average gaps between contiguous orders of hierarchical complexity are equal to the average difference in Rash item difficulty for the lower of the contiguous hierarchical complexity groups. There are 5 null hypotheses: $a_n = 0$, n = 7, 8, 9, 10, 11. The alternative hypothesis

Appendix B continues on the following page.

Appendix B continues from the previous page.

is that the average gap is bigger or smaller than the average difference in Rash item difficulty scores. There are 5 alternative hypotheses: $a_n \neq 0$, $n = 7, 8, 9, 10, 11$.

t-tests can be used to test the nulls against the alternative. The 5 t-tests are:

$t = a_n/\mathrm{SE}(a_n)$

The result of the tests show that

$t7(97) = -10.014$, $p < 2e - 16$ ***

$t8(97) = -9.667$, $p < 8.5e - 16$ ***

$t9(97) = -10.555$, $p < 2e - 16$ ***

$t10(97) = -10.848$, $p < 2e - 16$ ***

$t11(97) = -10.499$, $p < 2e - 16$ ***

All the tests show that the null hypothesis is rejected and the items breaks are significantly different from the gaps. In addition, because the average of Items Breaks at each order is smaller than the average of gaps, average item breaks are significantly smaller than the average gaps. Therefore, we have shown that gaps exist.

Appendix C
Lack of fit Test

The separate means model assigns a parameter to every group. The parameters represent the group means.

Rasch scaled item difficulty = $\beta_7 + \gamma_8 I_{8i} + \gamma_9 I_{9i} + \gamma_{10} I_{10i} + \gamma_{11} I_{11i} + \varepsilon_i$

I_{ni} is a dummy variable. i = 7, 8, 9, 10, 11. I_{ni} = 1 when the item is at the order of hierarchical complexity denoted by the subscript. I_{ni} = 0, when the item is at a different order of hierarchical complexity from the subscript.

After data is fitted to the model, it shows that $r = 0.988$, $r^2 = .977$, $F(4, 97) = 1023$, $p < 0.001$. In addition, residual standard error (RSE) = 0.363, $df = 97$

The Linear Regression Model maps a linear relationship between the independent and dependent variables.

Rasch scaled item difficulty = $\alpha + \beta * x$

x = Item order of hierarchical complexity,

α = intercept,

β is the slope.

Result shows that Residual standard error (RSE) = 0.3681 $df = 100$

Testing for lack of fit:

H_0: the linear regression model explains significantly less variance than the seperate means model.

H_1: the linear regression model and the seperate means model explains equal amount of variance in the data.

$$
\begin{aligned}
F\text{- Stat} &= [SSRes_{LR} - SSRes_{SM}]/[df_{LR} - df_{SM}]/\delta^2_{SM} \\
&= [RSE_{LR}{}^2 * df_{LR} - RSE_{SM}{}^2 * df_{SM}]/[df_{LR} - df_{SM}]/\delta^2_{SM} \\
&= (0.3681^2 * 100 - 0.363^2 * 97)/(100 - 97)/0.383^2 \\
&= 1.9438,\ df = 100 - 97 = 3 \\
p &= 0.1276
\end{aligned}
$$

The result shows that we cannot reject the null hypothesis that the linear regression model explains as much variance as the separate means model, providing supporting evidence to linearity.

Appendix D
Test for Equal Spacing

The following model is constructed to test for equal spacing.

$$\text{Rasch scaled item difficulty} = B_7 I_{7i} + B_8 I_{8i} + B_9 I_{9i} + B_{10} I_{10i} + B_{11} I_{11i} + \varepsilon_i \quad [1]$$

Rasch scaled item difficulty is the difficulty of item *i*.

I_{ni} is a dummy variable. $i = 7, 8, 9, 10, 11$. $I_{ni} = 1$ when the item is at the order of hierarchical complexity denoted by the subscript.

$I_{ni} = 0$, when the item is at a different order of hierarchical complexity from the subscript. Note that B stands for beta, the true underlying parameter.

β is its estimate.

Note also that I have added a subscript *i* denoting an observation.

Note that all of the items are in one of the five groups; hence there is a linear dependency among the indicators of every item *i*. Next, transformation to equation [1] is applied.

For all items *i*,

$$I_{7i} + I_{8i} + I_{9i} + I_{10i} + I_{11i} = 1 \quad [2]$$

Solve equation [2] for I_{7i} , which is equation [3]

$I_{7i} = 1 - I_{8i} + I_{9i} + I_{10i} + I_{11i}$, and substitute the solution for I_{7i} into equation [1], [3]

$$\text{Rasch scaled item difficulty} = \beta_7 (1 - I_{8i} + I_{9i} + I_{10i} + I_{11i}) + \beta_8 I_{8i} + \beta_9 I_{9i} + \beta_{10} I_{10i} + \beta_{11} I_{11i} + \varepsilon_i. \quad [4]$$

Combine similar terms,

$$\text{Rasch scaled item difficulty} = \beta_7 + (\beta_8 - \beta_7) I_{8i} + (\beta_9 - \beta_7) I_{9i} + (\beta_{10} - \beta_7) I_{10i} + (\beta_{11} - \beta_7) I_{11i} + \beta_i \quad [5]$$

Let

$\gamma_8 = (\beta_8 - \beta_7)$,

$\gamma_9 = (\beta_9 - \beta_7)$,

$\gamma_{10} = (\beta_{10} - \beta_7)$,

$\gamma_{11} = (\beta_{11} - \beta_7)$.

And substitute these values into equation [5],

$$\text{Rasch scaled item difficulty} = \beta_7 + \gamma_8 I_{8i} + \gamma_9 I_{9i} + \gamma_{10} I_{10i} + \gamma_{11} I_{11i} + \varepsilon_i \quad [6]$$

Appendix D continues on the following page.

Appendix D continues from the previous page.

Equation [6] describes a linear regression producing least-squares estimates $\{\beta_7, \gamma_8, \gamma_9, \gamma_{10}, \gamma_{11}\}$ for $\{B_7, \Gamma_8, \Gamma_9, \Gamma_{10}, \Gamma_{11}\}$.

β_7 = the average value of the Rasch scaled item difficulty for items in order 7.

γ_8 = is the estimate of the difference between the average Rasch hierarchical order 8 score and the Rasch hierarchical order 7 score.

Accordingly, $\beta_7 + \gamma_8$ estimates the average Rasch scaled item difficulty at order 8. Similarly, $\{\gamma_9, \gamma_{10}, \gamma_{11}\}$ estimate the difference between the average Rasch scaled item difficulty at orders {9, 10, 11} and the average Rasch scaled item difficulty at order 7. Accordingly, $\{\beta_7 + \gamma_9, \beta_7 + \gamma_{10}, \beta_7 + \gamma_{11}\}$ estimate the average Rasch scaled item difficulty for items in hierarchical orders {9, 10, 11}.

Spacing is defined as the increment from the average of Rasch Scaled Item Difficulties of a lower order to the average of Rasch Scaled Item Difficulties of the next higher order. The hypothesis we are fundamentally interested in testing is whether there is the same incremental difference between $B_7, B_8, \ldots,$ and B_{11}. The null hypothesis for this analysis is that the spacings between each pair of adjacent orders are the same. There are three versions of this null hypothesis, as indicated in the following.

H_{01}: The Rasch scaled item difficulty difference between order 9 and 8 is the same as the Rasch scaled item difficulty difference between order 8 and 7.

H_{02}: The spacing between order 10 and 9 is the same as the spacing between order 9 and 8.

H_{03}: The spacing between order 11 and 10 is the same as the spacing between order 10 and 9.

First, consider H_{01}. The spacing, or Rasch scaled item difficulty difference between order 9 and 8 is estimated by

$\beta_9 - \beta_8 = (\beta_9 - \beta_7) - (\beta_8 - \beta_7) = \gamma_9 - \gamma_8$

The spacing between order 8 and 7 is estimated by $(\beta_8 - \beta_7) = \gamma_8$. Therefore,

H_{01}: $(\gamma_9 - \gamma_8) = \gamma_8$,or, $\gamma_9 - 2\gamma_8 = 0$

Similarly,

H_{02}: $\gamma_{10} - 2\gamma_9 + \gamma_8 = 0$

H_{03}: $\gamma_{11} - 2\gamma_{10} + \gamma_9$

A one-sample *t*-test can be conducted to test these hypotheses. The formula is

t = (Estimate of Spacing 1 − Estimate of Spacing 2 − 0)/SE((Estimate of Spacing1 – Estimate of Spacing2)

For H_{01}: $\gamma_9 - 2\gamma_8 = 0$,

$SE(\gamma_9 - 2\gamma_8) = (Var(\gamma_9) + 4\,Var(\gamma_8) + 4\,Cov(\gamma_9, \gamma_8))^{1/2} = 0.290$

Hence the *t* value with 97 degrees of freedom is

$t(97) = ((\gamma_9 - 2\gamma_8) - 0)/\,SE(\gamma_9, \gamma_8) = 0.240$, $p = 0.595$

Therefore, we cannot reject the null hypothesis that the spacing between order 9 and 8 is the same as the spacing between order 8 and 7.

Appendix D continues on the following page.

Appendix D continues from the previous page.

Next, for H_{02}, $\gamma_{10} - 2\gamma_9 + \gamma_8 = 0$

SE ($\gamma_{10} - 2\gamma_9 + \gamma_8$) = (Var(γ_{10}) + 4 Var(γ_{19}) + Var(γ_8) + 4 *Cov* (γ_{10}, γ_9) + 2 *Cov* (γ_{10}, γ_8) + 4 *Cov* (γ_9 , γ_8))$^{1/2}$ = 0.3864347

Hence the *t* value with 97 degrees of freedom is

$t(97) = (\gamma_{10} - 2\gamma_9 + \gamma_8) - 0)/ SE(\gamma_{10} - 2\gamma_9 + \gamma_8) = 0.0526$, $p = 0.479$

Therefore, we cannot reject the null hypothesis that the spacing between order 10 and 9 is the same as the spacing between order 9 and 8.

Lastly, for H_{03}: $\gamma_{11} - 2\gamma_{10} + \gamma_9 = 0$

SE($\gamma_{11} - 2\gamma_{10} + \gamma_9$) = (Var($\gamma_{11}$) + 4Var ($\gamma_{10}$) + Var($\gamma_9$) + 4 *Cov* ($\gamma_{11}$, γ_{10}) + 2 *Cov* (γ_{11}, γ_9) + 4*Cov* (γ_{10}, γ_9))$^{1/2}$ = 0.3778

Hence the *t* value with 97 degrees of freedom is

$t(97) = (\gamma_{11} - 2\gamma_{10} + \gamma_9) - 0)/ SE(\gamma_{11} - 2\gamma_{10} + \gamma_9) = 0.7949$, $p = 0.214$

Therefore, we cannot reject the hypothesis that the spacing between order 11 and 10 is the same as the spacing between order 10 and 9.