Behavioral Development Bulletin
2016, Vol. 21, No. 2, 223–239

1942-0722/16/$12.00 http://dx.doi.org/10.1037/bdb0000039

Comparing Different Ways of Using the Model of Hierarchical Complexity to Evaluate Graduate Students

Patrice Marie Miller and Darlene Crone-Todd
Salem State University

University educators at all levels, including undergraduate and graduate, are increasingly being asked to assess their efforts to show that students are learning skills needed for work and life. The question continues to be how to most effectively assess student competencies. Here, we introduce the model of hierarchical complexity as a candidate for such assessment. The primary reason that such a model could be useful is that it postulates that the development of competency results from the mastery of tasks that occur in a sequence from least to most hierarchically complex. We use the model to assess the competence level of students entering or already enrolled in a master's program in counseling, and predict students' grades in that program. Two separate studies were carried out. In the first study, narrative statements that students wrote in order to be admitted to a master's program in counseling were coded using a method developed from the model of hierarchical complexity. Statements in students' essays were coded at 5 stages, from Concrete to Metasystematic. The mean stage found was 10.76, or between the Abstract stage and the Formal stage. Although those students with slightly higher mean scores also had slightly higher GPAs, with the small number of coded participants and the very restricted range of student GPAs, we were not able to show that the difference was significant. In the second study, we used a structured method of assessing participants' stage, and contrasted that with the coding of a short narrative statement collected at the same time. In this second study, the scores from the coding of the narrative statement did significantly predict GPA, but not the scores from the structured instrument. The results of the 2 studies provide some support for the idea that using the model of hierarchical complexity to assess and predict graduate school success would be useful, and should encourage those who have interests in this area to carry out additional studies with larger and more variable samples as well as a variety of related instruments.

Keywords: model of hierarchical complexity, graduate school, therapist, development

It is common in higher education to discuss the question of what developmental competencies are needed now and in the future. In general, and especially for undergraduate students, such discussions might focus on such skills as critical thinking. In training graduate students, the skills to be emphasized extend beyond those developed at the undergraduate level, and also depend more upon the kind of specialization for which they are training. Because competencies in different training areas will be different, within the present article we will restrict the focus to studies of graduate students being trained in counseling psychology at the master's level.

Currently, in the United States, people who work in therapeutic settings may be trained and licensed at different levels and in different disciplines. For example, clinical or counseling psychologists are typically trained and licensed at the doctoral level (e.g., PhD or PsyD). At the master's level, individuals may be trained and licensed to become mental health counselors or marriage and family therapists. Some individuals trained and licensed as social workers or as behavior therapists also provide therapeutic services. This article focuses only on those individuals who are training at the master's level to

Patrice Marie Miller and Darlene Crone-Todd, Department of Psychology, Salem State University.

Correspondence concerning this article should be addressed to Patrice Marie Miller, Department of Psychology, Salem State University, 352 Lafayette Street, Salem, MA 01970. E-mail: pmiller@salemstate.edu

become mental health counselors or marriage and family therapists.

With this population in mind, we start by asking "What kinds of competencies might be necessary for an individual to effectively function as a master's level therapist?" We then examine the usefulness of a general model of "smarts" or of competence, called the *model of hierarchical complexity* (MHC), for predicting how well students do in a master's level counseling program.

The field of counseling takes an approach that tends to emphasize competence in specific areas. For example, the Council of Counseling Psychology Training Programs (CCPTP) proposed a set of counseling competencies for programs to follow (see http://www.ccptp.org). This set of competencies has two important features. First, there are a series of domains in which students should become competent, such as professional values and attitudes, individual and cultural diversity, ethical/legal standards and policy, reflective practice/self-assessment/self-care, and relationships, to name a few. The second feature of these competencies is that they contain a developmental component, in that students may be rated as completely lacking, emergent, or proficient, in a given area.

Clearly, becoming proficient in domains relevant to counseling is an important goal, and thus an important aspect to be assessed. However, there are a few possible problems with such an approach. First, although these lists of competencies do include specific behaviors that would occur at the "emergent" level or at the "proficient" level, they basically rely on a matching technique. If students show the specific behaviors listed, they would then be categorized as meeting the standard. If they do not show these behaviors, they will typically be judged as not meeting the standard. Although often this is a workable, objective procedure, it is also possible that the student is proficient, but that some of the specific behaviors in which they engage to demonstrate that proficiency are not listed. So, raters might have to make judgments on behaviors that are not included in the lists of competencies. In other words, it is not clear what dimensions underline the judgments being made. This means that the definitions for the competencies may not be completely accurate or valid, even if they are reliable.

A second, and more important, issue is that the competencies are essentially behavioral outcomes of training. This conception does not address the process by which students attain the outcomes. It is likely that given the exact same training conditions, some students will attain the required level of competency and others will not. The question we pursue in the present article is whether there are individual factors at the outset (i.e., when applying to or during training) that may make it more or less likely that students will attain a given level of competency.

The approach we explore here involves examining the complexity with which students appear to conceptualize constructs about themselves and others. The article also relates that complexity to some possible outcomes within a graduate program. In order to accomplish this, we rely on the MHC, which allows for the scaling of the difficulty of a variety of tasks. The MHC will be described next.

The Model of Hierarchical Complexity

The MHC is a neo-Piagetian and quantitative behavioral-developmental theory that can be used to analyze the difficulty or complexity of tasks (Commons & Richards, 1984; Commons, Trudeau, Stein, Richards, & Krause, 1998). The model is based on the assumption that a large number of tasks exist in the environments of all animals, whether humans or not, and that these tasks occur in sequences that can be ordered as to their difficulty. The model postulates that these sequences exist in different *domains* of behavior including problem solving, personal, social and others. The primary reason that such a model could be useful is that it postulates that the development of competency results from the mastery of tasks that occur in a sequence from least to most hierarchically complex.

What the model allows for is a scaling of the difficulty of tasks, which is operationalized in terms of a measure called the order of hierarchical complexity (OHC). Tasks at each higher order result from the combination and coordination of at least two actions from the next lower order task. Figure 1 illustrates this property of the model. The OHCs, with examples at each order, are shown in Table 1. Note that when an organism successfully completes a task at a particular order, they are then said to be

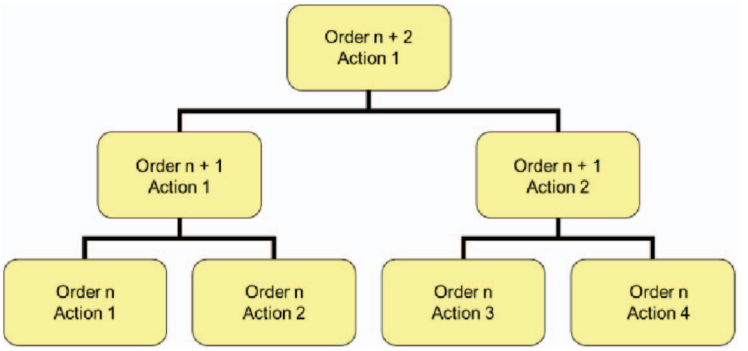

Figure 1. Tree diagram for Model of Hierarchical Complexity.

performing at the stage that has the same number as the order of the task. For example, if an organism successfully completes a task at Order 10, then they are said to perform or behave at Stage 10 on that task.

Note that this model is notable for describing orders and stages of performance that are beyond Inhelder and Piaget's (1958) Formal stage. Only the orders up to Metasystematic (Order 13) are shown, as these are most relevant for the current article. Additional postformal orders (14, 15, 16) are described in Commons and Ross (2008) and in Ross, Commons, Li, Stalner, and Barker (2014). Examples of the stage performances on the tasks studied in this article will be provided in the Method section of each Study.

In short, this model provides an underlying mechanism for stage changes across development. Note that the stages map relatively well onto other conceptions of stage development, for stages from the Formal, on down, especially for those conceptions that included half stages as stages (e.g., Fischer, 1980; Pascual-Leone, 1970).

Because this stage theory includes postformal stages, it is particularly important for understanding the performances of graduate students, because in earlier theories (e.g., Inhelder & Piaget, 1958) development was essentially seen as reaching its pinnacle during adolescence.

There have been some previous attempts to use developmental factors, including stages, to evaluate counseling or clinical competencies. For example, Eriksen and McAuliffe (2006) studied entering counseling students attending three different universities, finding that the best predictor of their measure of counseling skills was a measure of moral development (Rest's Defining Issues Test; developed by Rest, Narvaez, Bebeau, & Thoma, 1999). It should be noted that the performance of the students on this test was at the "conventional" level. According to Rest et al. (1999) this is equivalent to the Kohlberg's Stage IV, which according to these authors is generally seen as equivalent to the Formal Operational Stage. This is a stage of reasoning in which people are primarily interested in maintaining current norms, not moving beyond them. Eriksen and McAulliffe argue that this fits with their experience, when training counseling students, of having to encourage them to move beyond their existing frameworks to consider the frameworks of others. Because these were beginning students these results are not surprising. It would be useful to have data from counseling students at a variety of levels, not just those in the initial counseling theories courses, as this would allow one to evaluate incoming, progressing, and graduated students in terms of their development of competencies, moral development, and complex thinking.

A more recent developmental conception of counseling and psychotherapy has been presented by Basseches and Mascolo (2010). Although this very detailed book cannot be easily distilled in just a few sentences, one theme seems important. This theme is that psychotherapy is a developmental process in which areas of conflict are resolved by organizing and reorganizing the elements that are in conflict so as to form more differentiated, inte-

Table 1
Orders of Hierarchical Complexity, From Order 0 (Calculatory) to Order 13 (Metasystematic) Along With Examples at Each Order

Order name	Order no.	Example
Calculatory	0	Follow a programmed set of instructions. Example: computer program.
Automatic	1	Tropisms, sensitization, habituation, unconditionable reflexes. Example: Paramecium moves away from light (Mingee, 2013).
Sensory or motor	2	Respondent conditioning. Example: On hearing mother's voice, infant turns head in that direction, begins rooting.
Circular sensory–motor	3	Operant conditioning. Example: When infant babbling is followed by vocalizing and smiling from adult, infant babbles more.
Sensory–motor	4	Forms concepts. Example: Animals from a variety of species learn discriminations of concepts, such as same/different.
Nominal	5	Relates two (or more) concepts, including relating a concept to its name. Example: can say the word "same" or name other concepts, such as "boy."
Sentential	6	Combines names into short sequences or sentences. Example: A child says the names of a few numbers or letters, in order. Says short sentences.
Preoperational	7	Combines sentences into sequences. As a result, makes simple deductions, follows a list of sequential acts. Does not relate these to reality. Example: Tells a story of a few sentences.
Primary	8	Relates single actions to reality. Applies simple deduction and empirical rules. Recounts what has happened reasonably accurately. Understands their own perspective, or that of another person, but does not relate these two at the same time. Example: "I was sad because I lost my toy."
Concrete	9	Simple logical deduction and time sequences are used to describe actual instances. The instances are actual because they occur in past or present time. They are composed of specific things, incidents events, actions, actors and places. Coordinates own perspective with that of one other (at a time). Example: "When my grandmother was sick, I took care of her."
Abstract	10	Classes, variables, or "variabilized" (abstract) concepts are formed from concrete instances. Example: "I felt really bad that day."
Formal	11	Relates one variable to another variable. Example: "When people don't respond to me quickly, I tend to feel insecure."
Systematic	12	Coordinates more than one variable as input, constructs multivariate systems and matrices, situate events and ideas in a larger context. Example: "Both how the parent behaves, and how the child responds will affect whether the child learns this concept."
Metasystematic	13	People act on systems of relations from the previous order. Metasystematic actions compare, contrast, transform, and synthesize systems. Example: "Children serve different functions in different kinds of societies. In agrarian societies, they may function primarily as field workers. But in today's industrialized societies, they serve an important function as consumers."

grated and adaptive wholes (see, e.g., p. 22). A great portion of this book is, as a result, prescriptive in that it is advocating a particular approach to psychotherapy. Progress in psychotherapy is seen as occurring through a dialectical process that includes both therapist and client, who "co-act" to reorganize the conflicting elements. This process is illustrated through the use of case studies. Although the information in the book could possibly be used to differentiate more versus less experienced counselors and psychotherapists that was not a primary goal of the book. What is useful about these ideas from the point of view of this article is that it argues strongly for the importance of a developmental view in understanding counselors and counseling.

In this article, we take a somewhat simpler view of what underlies counseling skills. Our view is that one can conceptualize counseling (or psycho-

therapy) skills in terms of the hierarchical complexity of the tasks that the counselors-in-training (and ultimately the counselors and the psychotherapists) are undertaking. Therefore, the MHC measures should be related to outcomes in a counseling program.

At the current time, there is not very much research on using the MHC to determine the stage at which students complete academic work. One study by Crone-Todd and Gonsalves (2010) found that undergraduate Honors students in psychology wrote at a Formal stage in their thesis projects, whereas nonhonors student writing tended to use more Abstract-stage writing. The reasons for this difference in complexity of writing could be due to many factors, including academic preparedness, feedback from committee members, and so forth This suggests that using such a measure could be useful as a way to measure individual differences, and that these differences could be then related to measures of competency, such as whether a student is in an honors program or not. Another study (Kjellström, Ross, & Fridlund, 2010) examined the stages of ethical reasoning in nurses studying in a Ph.D. program. This study also found that few nurses included arguments that were scored above the Formal stage.

In the first study we used the MHC to code the real world behavior of students, specifically their writing. The rationale for this is that we often need information about the existing behavior of workers or students. In cases in which we have real-world behavior, the model can be used to score a variety of kinds of behaviors. This study will answer the question as to what stages are typical of such students. Based on the previous study of Crone-Todd and Gonsalves (2010) we would expect to see graduate school applicants writing at the Formal stage or above. We planned also to relate the stage of students' writing to their grades in the Counseling program.

Method—Study 1

Participants

The study used archival data from 14 students who either currently, or in the past, were enrolled in a master's level counseling program at Salem State University. In the Commonwealth of Massachusetts, such programs train individuals who may then apply to become licensed as mental health counselors, or marriage and family therapists (depending upon the specific program they were enrolled in). In order to become licensed, students must complete a minimum of 60 credits of graduate course work, including a 12-credit counseling-training sequence called the Practicum and Internship. Content courses are 3 credits each.

The procedure for obtaining the specific participants was as follows. First, all 196 archived student files in the program were numbered. Students who had files had been admitted to the program, although not all continued to be enrolled. From this original pool of 196, a random sample of 100 files was selected. The coding was done by two Professors, who had considerable experience using the model to either code statements or to create instruments, and by two students. Because the students were learning to code, coding each participant's narrative statement took a week to prepare (with everyone coding it on their own), and then one or two meetings after that for the group to discuss. As a result only 14 student statements were coded during the time period in which all four coders were available. These 14 individuals were also chosen randomly from the overall pool of 100. A master list of the 100 students was kept to allow matching of a student's number to an I.D. and name, should that be necessary. Student names were removed from all narrative statements before coding took place. It is important to add that one of the professors had no knowledge at all about any of the students. The second professor had read all 196 narrative statements at some time, when those students were originally admitted, but had no direct knowledge of the students' classroom performance (she had never had any of them in a course). The students, who were advanced undergraduates, were in a different program from the participants and so would not have known them.

Applicants to the counseling program were from a wide range of institutions, with only a minority of applicants having attended the same university for their undergraduate study. Only about one third of the students admitted were applying directly from their undergraduate program. Other applicants had already been working in the field, or had been working in a different field and were career changers. Because this is a small sample of the overall students, it may or may not be representative of the larger group. At this point, the purpose of the study was to use a sample to test the feasibility of

coding such essays using MHC, and also whether there was evidence that results of this coding were meaningfully related to outcomes they would be expected to be related to.

The essays to be coded were kept in student files that were kept at the School of Graduate Studies. As such, they were judged by the institutional review board to be archival data. Because of its use of archival data, the study was judged to be an exempt study. In any case, the confidentiality of student records is protected, and no information will be presented here that could identify the students whose essays were coded.

Materials

The archived materials to be coded were students' Statement of Purpose, written as part of their applications for admission to the master's program in counseling. These essays were required to be approximately 500- to 1,000-words. The specific questions to which students are asked to respond are: "What factors in your personal and professional history have influenced you in the past and now lead you to seek admission to this program at Salem State University? What are your long-term professional goals, and how will this program help you meet them? Is there any other information that you think would help the admissions committee to evaluate your application?"

Procedure

Each essay was scored by four coders. Two of the coders were experienced with previous scoring using the MHC, and the other two coders (who were students) were learning both the model and how to code as they participated as part of the research team. Coding involved reading each sentence in an essay individually, and comparing it to the definitions of the various orders of complexity in the MHC. If possible, coders also noted the transition step of the sentence. Because the particular transition steps are not attended to in the analysis, this will not be discussed further here. Once all coders completed their scoring of an essay, they then met together and compared their coding for each sentence. When differences between coders did exist, the sentence and its coding was discussed at greater length, until the entire group agreed upon what order to assign it to. Note that Commons and Jiang (2014) updated the model, adding one additional order and stage at the beginning of the sequence. All Order and Stage numbers used here conform to the new numbering.

An example for each order of complexity is provided next.

Examples of Narration at Each Stage

Concrete stage (completely completes a task at order 9). *But at that time, my grandfather, who I had lived with for years, was diagnosed with cancer* [Participant 170].

Tells a story with concrete facts, but one that at least implicitly coordinates two individuals' behavior.

Abstract stage (completes a task at order 10). *This is a difficult transition to make, from high school to college* [Participant 4].

The difficulty of the transition names a variable called degree of difficulty.

Formal stage (completes a task at order 11). *In these positions, I received some personal satisfaction knowing that I was helping put quality textbooks into the hands of students* [Participant 49]

If I help put quality textbooks in the hand of students then I received some personal satisfaction. This relates two Abstract stage variables:

X: textbooks differ in quality [in this case, the individual was putting "quality" textbooks in the hands of students]

Y: personal satisfaction differs

X: – > Y

Systematic stage (completes a task at systematic order 12). *Yet, while the faces are different, their stories have a common theme: past trauma, neglect and/or abuse, loss, diagnosed mental disorders, all of which lead up to them using the substance(s) to "self-medicate," to stop or numb the feelings, or just fill the void that has been left inside of them* [Participant 173].

Past trauma, neglect and/or abuse, loss, diagnosed mental disorders are multiple causes of them using the substance(s) to "self-medicate", and/or to "fill the void" that has been left inside

(two outcomes). So, in this case, multiple causes are related to one or more outcomes.

Transitional to metasystematic (gives some indication that might complete a task at order 13). *These experiences have allowed me to witness the family dynamics in two entirely different cultures that I can share through discussions with fellow classmates and faculty.*

This student grew up in a country in Eastern Europe, during the time that it was still part of the former Soviet republic. Earlier in this paragraph, and earlier in the essay, she discusses the topic of "family dynamics" in this context. Then, she mentions that having lived with and worked for two American families, she later on experienced "family dynamics" within families in this different culture. Earlier she describes "family dynamics" as a system containing multiple inputs and multiple outputs. Given the groundwork that was set up earlier in this essay we coded this as Metasystematic. Note that the two systems are not explicitly intercoordinated; so it is more properly coded as Systematic, in the transition step of "smash" on the way to Metasystematic.

Results—Study 1

The data for each individual consisted of a distribution of how many sentences were coded at each order of complexity. It is not yet clear as to what aspect of participants' performance is most important to focus on, because little MHC-scored data has been published. Should analysis focus only on mean stage scores, for example? Or, should it look at the highest score obtained, and how many sentences with this highest score there were? We will take the strategy here of first presenting the data with respect to how sentences were scored in the students' essays in some detail. This will be done in order to illustrate what different ways of analyzing the information show. Second, we will relate the patterns seen to student grades.

First, we show the distributions of stage scores for each of the participants (see Figure 2). These are shown for all the participants, as this is the first study to publish data of this kind. As can be seen from the distributions, there is considerable variability between individuals in the coded stage of sentences that were included in their narrations. Some individuals, such as # 4, #77, #122 show sentences across the range of Orders, from Concrete up to at least Systematic or slightly above. Concrete sentences in these narratives are often used to provide specific facts about an event.

A few individuals, such as #49, #99, #121, #152, #157, show a preponderance of sentences that are categorized as being either between Abstract and Formal (more typical perhaps of early adolescence), or between Formal and Systematic (more typical of undergraduate students). Across the 14 participants, the graphs show that 5 wrote statements that contained 10% or more sentences at Systematic and Metasystematic. The remainder of the participants either showed no sentences at Systematic or above (# 152, #171, # SRAD) or included less than 5% of such sentences at these higher orders (#17, #77, #99, #121, #157, #170). For most essays, 5% would be the equivalent of one sentence.

We next generated a number of descriptive statistics to help summarize the information in the graphs. First, we looked at essay length, in terms of number of sentences. The mean length of an essay was $M = 29.21$ ($SD = 9.72$), with the Median = 27.5 sentences. Lengths ranged from 18 to 59. The length was negatively correlated with mean stage (-.388), suggesting that longer essays had lower mean stage of sentences, however this was not significant.

An overall mean stage for each essay was calculated by summing the stage numbers assigned to each sentence, and then dividing that by the number of sentences. The range of the individual means was from 10.17 to 11.08, or ranging from the Abstract to the Formal stage. Across participants, the overall mean stage (so the mean of the means) of the sentences was 10.76 ($SD = .27$), or between the Abstract and the Formal stages. The Median Stage was 11, or Formal Stage, with the range of the individual Medians running from 10 to 11.25. The minimum stage found in this scoring was 9, or Concrete, with the range of the individual minimum stages running from 9 to 10. The maximum stage was 13 (Metasystematic), with the median of that maximum being 12, or Systematic, and the range of maximum stages being 11.5 to 13.

Another way to examine these data is to look at the proportion of "higher stage" sentences in each essay. In this case, we define "higher stage" as being at least in transition between Formal and Systematic stages. These findings are presented in Figure 3. As can be seen, between 0.31 and 0.80 of the sentences in the 14 essays were between the

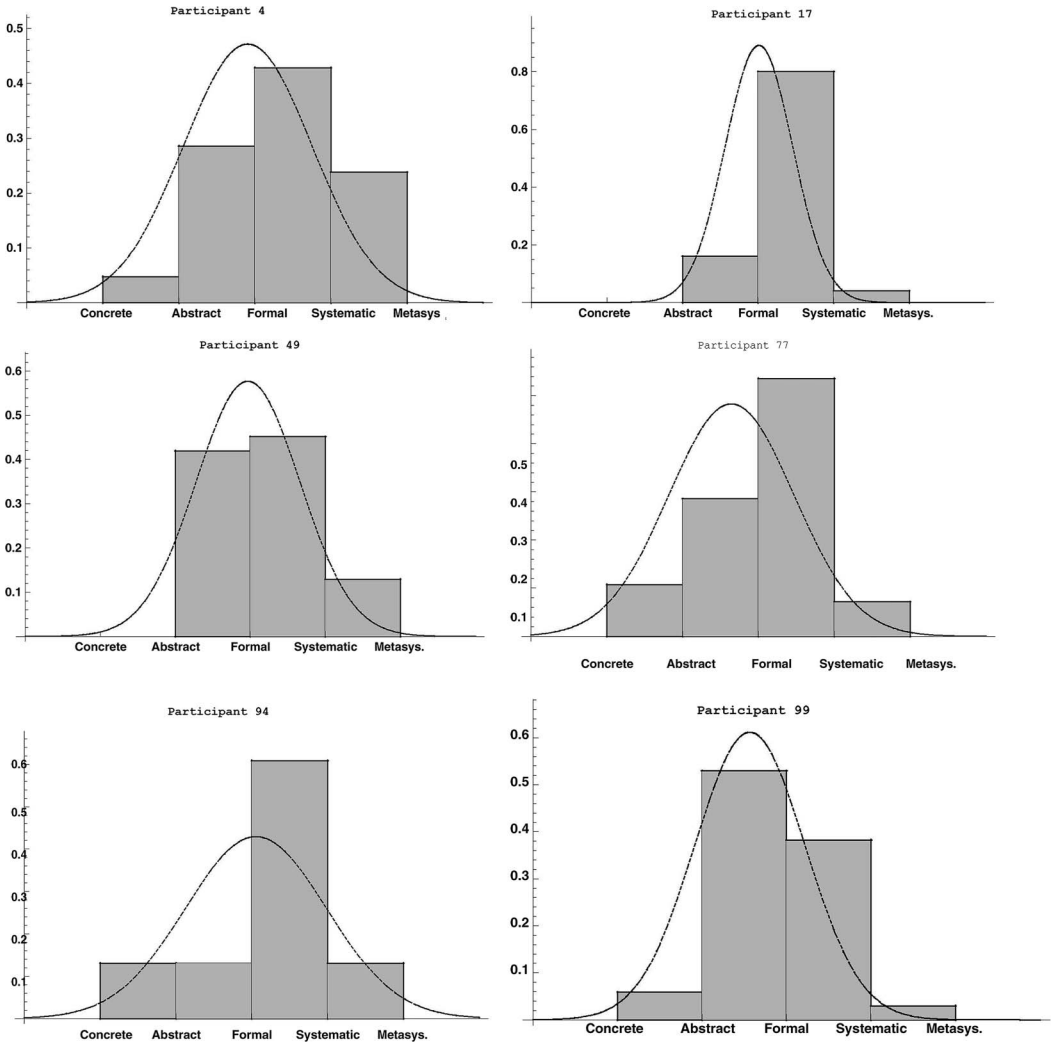

Figure 2. Distributions of statements at different stages for each participant.

Formal and Systematic (with overall $M = .495$, $SD = .12$). In fact, 12 of the participants used sentences between the Formal and Systematic stage at least 40% of the time in their essays, with two using them less frequently. When we examined essays for the proportion of sentences between Systematic and Metasystematic, we found that three individuals showed no fully Systematic stage sentences. Of the remaining 11, the proportions ranged from .01 of the sentences to 0.27 (excluding the zero cases, the $M = .108$; and $SD = .09$). Because there was only one individual who had a sentence scored as Metasystematic, this proportion was added into the previous category.

Relationship of Stage Levels to Participant Grades

In this analysis we will examine the relationship of mean stage to both overall student grade point averages (GPA) and to student average grade across two courses considered the most difficult in the program. For this analysis, grades from 12 of the 14 students were available. Overall GPA for these graduate students ranged from 3.355 to 3.955, a very restricted range. We found that overall GPA was positively correlated with Mean Stage, $r(12) = .318$, however, this was not significant. In a second analysis, we divided the partic-

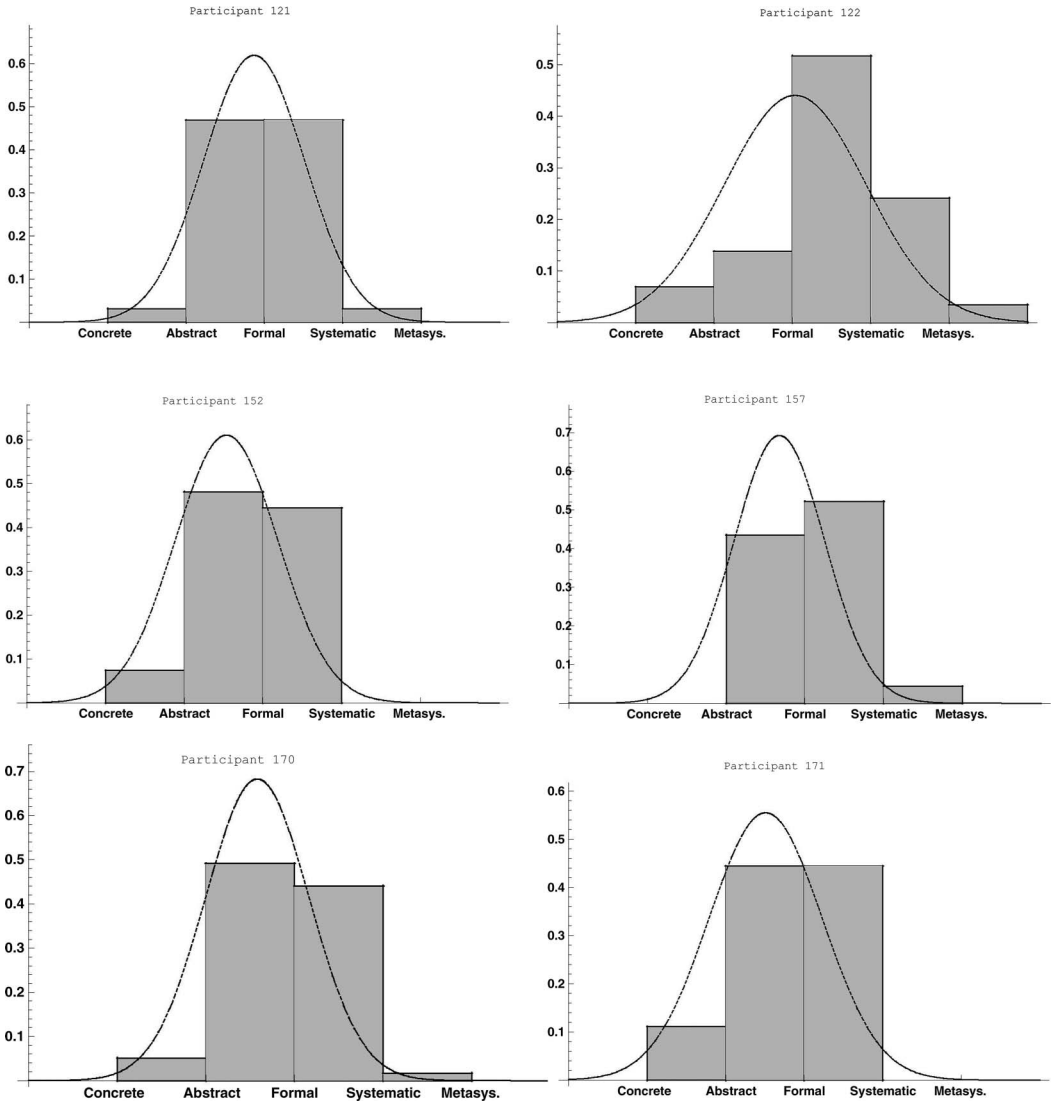

Figure 2 (continued).

ipants' mean stage into two groups, those whose mean stage was equal to or less than 11 (Formal; $n = 8$) and those whose mean stage was greater than 11 ($n = 4$). This analysis showed a mean difference in the expected direction. That is, individuals whose average score was less than Formal had a mean GPA of 3.73 ($SD = .225$), whereas those whose average score was greater than Formal, had an average GPA of 3.91 ($SD = .083$). The analysis of GPA in the two hard courses, showed a similar pattern, but was even weaker.

Discussion—Study 1

We had hypothesized that students going on to graduate school should be likely to show at least Formal operational reasoning, and perhaps some Systematic stage reasoning. The data show that participants' mean stage demonstrated in their written essays was 10.76, or in between the Abstract and Formal stages. The range of the individual means was 10.17 to 11.08. In other words, no participant obtained a

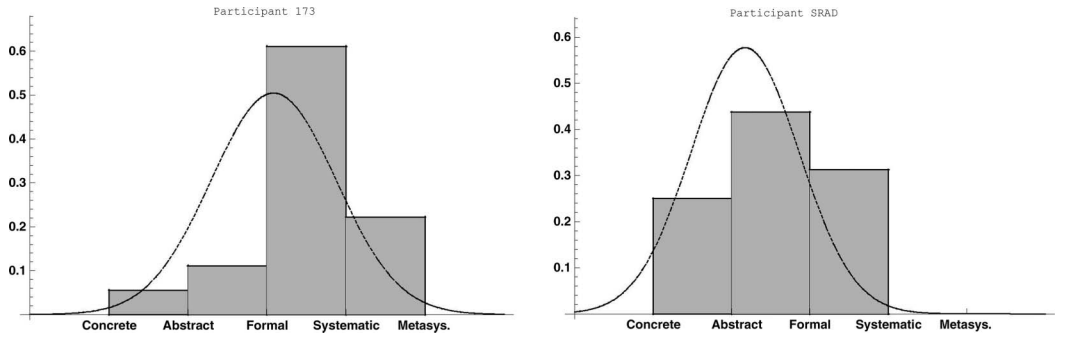

Figure 2 (continued).

mean higher than Stage 11, or Formal stage. This should not be of great concern with respect to reading such statements of purpose. In reading the essays, it seemed to the coders as if some number of Concrete and Abstract statements were inevitable, because they are important in recounting specific facts of one's background and also including reactions to those facts, which are often stated in Abstract Stage terms (e.g., as in "This is a difficult transition . . .").

It was for this reason that we next looked at the proportions of the different stages within each essay. We found that most participants used sentences between Formal and Systematic stage at least 40% of the time. Many fewer of the sentences overall were between Systematic and Metasystematic (the mean percent of sentences between these two stages was around 10%). Only two individuals had at least 20% of their sentences categorized as at least Systematic. These findings do show that the majority of the participants would be capable of learning Systematic content in graduate courses, as they already exhibit use of it within their own work.

We found that stage was positively related to participants' GPA, however, with the small sample size and the very small differences between participants' GPA's, we were not able to determine if there was a significant relationship.

Because there are few comparable validation studies at this time, it is difficult to say whether these findings are what is to be expected of incoming master's level students, or whether their essays were scored lower than would be expected. For many of the incoming graduate students in this study, the preponderance of sentences were scored at between Formal and Systematic. These findings are consistent with the findings of Crone-Todd and Gonsalves (2010), who showed that Psychology Honors' students generally wrote at the Formal stage. We would assume that if such students applied to graduate school, that their writing might be scored similarly to the students studied here. They are also similar to the findings of Kjellström et al. (2010). Even though their participants were in a doctoral program in nursing, those students' writings on the subject of ethics consisted primarily of statements at the Formal stage. They also found that 10% of nurses responses were at the Systematic stage, which is the same proportion found here.

The question is whether individuals who have at least a 4-year university education should engage in a higher proportion of Systematic stage thinking. Although Leite (2016) estimates that roughly 6% of the world population should be Systematic, a subgroup of the population that has both completed a 4-year university degree, and is enrolled in a graduate program should perhaps have a larger proportion of such higher stage thinking. One might also think that people would put considerable thought and effort into writing an admissions essay, and that the "demands" of such an essay require something closer to Systematic stage thinking. That is, you are being asked, at least implicitly to discuss different experiences from your past, and to combine them into a coherent story (a system?) that shows why your interests have developed in a certain way. Another possibility is that the use of an open-ended instrument, such as writing an essay, may make it more difficult for participants to show their competence to the highest degree. Because writ-

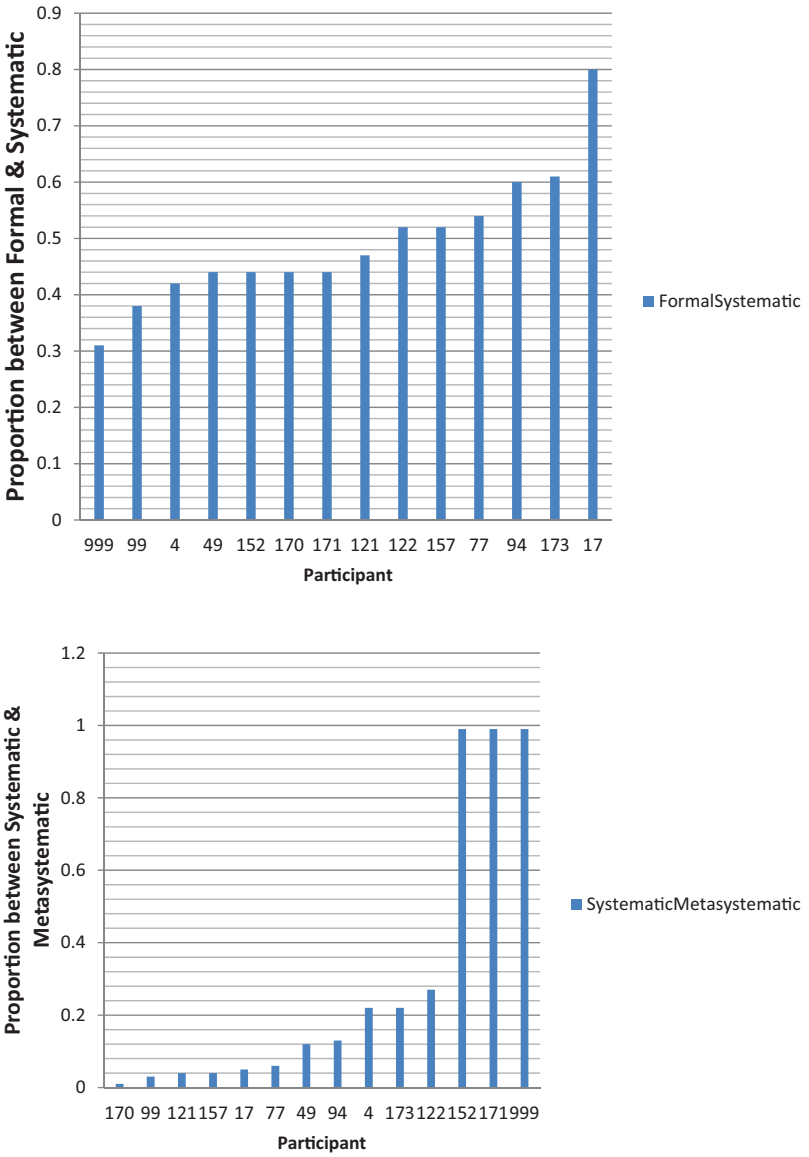

Figure 3. Frequencies of formal versus systematic sentences.

ers must both conceive of the issues at a certain complexity, and they must also have the writing skills to express their ideas at a certain complexity, an essay task no matter where or how administered may in effect underestimate the stage of tasks that participants are able to successfully complete. In order to see whether we could confirm some of the findings of this first study, we decided to undertake a second study. In this second study we used a structured instrument to measure stage, comparing it to the scoring of one open ended statement in the data. Because in the second study we were not also training students to score, and the entire instrument could be completed online, we were able to increase the overall sample size so that we will be able to ask some of the same questions, but perhaps come to a more definitive answer.

Method—Study 2

A number of recent studies by Commons and colleagues have created and studied tasks made up of problems or vignettes (e.g., Commons et al., 2006; Li et al., 2014). In the vignette studies, such as those of the counselor–patient relationship, these are usually constructed so as to have 5 or 6 subtasks that vary from the Preoperational, Primary or Concrete to the Metasystematic order. As the above studies have shown, this has been a powerful way to find confirmatory evidence for the Orders of Hierarchical Complexity.

Because the counselor–patient instrument can be relatively easily administered and scored, in this second study we can both determine the stage of performance of the participants and we can also more easily collect data on more participants. In addition, once we have their stage scores, we can more easily examine outcomes such as students' grades. That will add to the overall or construct validity of this instrument.

Participants

As in Study 1, participants were enrolled in a Master's level counseling program at Salem State University. An announcement was sent out to the student listserv, inviting all students to participate in the study. There were 35 participants who completed the entire survey, and so were eligible to have their data analyzed. Students varied in terms of the numbers of credits they had taken. None of the participants from the first study participated in the second study.

The major difference between the participants in Study 1 and those in Study 2 is that in Study 1, the participant data that were being coded were based on narrative statements written before the students began the graduate program. Here, the participants filled out the survey instruments at some point after they became matriculated in the program. The program usually takes at least two years to complete, and students who completed the instruments did so at various points within the program.

Measure

The counselor–patient survey was used in this study (Commons et al., 2006). This instrument consists of six vignettes. In each vignette, a different counselor informs their patient about a treatment and how they obtain consent for the treatment. Each vignette was designed to reflect a different OHC, ranging from Primary to Metasystematic. Previous studies have used Rasch analysis to show that participants do scale the vignettes in the manner in which they were designed (Commons et al., 2006). Examples of vignettes at two different orders are shown below. Examples of other vignettes can be found in the Commons et al. (2006) study. Note that when presented to participants the vignettes do not begin with the Order of Complexity labels shown here. Vignettes are also presented in a random order, rather than in sequence.

Concrete order vignette. Counselor Mason offers the patient a treatment preferred by colleagues. Mason says that others who are friends use this treatment. A colleague is called in to tell the patient again about the treatment. With great concern, Mason asks if the patient would like to hear a third person explain the treatment. Mason's patient is told that these people had good results with that treatment. Mason instructs the patient to support the treatment. Mason's patient thinks seriously about what Mason has said. Feeling that Mason knows best, Mason's patient prepares to undergo the treatment.

Systematic order vignette. Counselor Ellis offers a treatment which performs relatively better than others. Ellis relates the effects and side effects of each treatment including taking no action. Then Ellis asks the patient questions about the treatments making sure the patient understands. Ellis asks if the patient feels comfortable making a decision with the present information. Because the patient is satisfied, Ellis asks the patient to think carefully before choosing a treatment. Ellis asks the patient to think about what they have both said about the alternatives and then think about choosing. Feeling that Ellis knows best, Ellis's patient prepares to undergo the treatment.

After reading each Vignette, participants were asked to rate the methods of informing and obtaining consent in terms of seven dimensions or aspects of the informing and consenting. Two of the rating questions are included in this article: (a) Rate Counselor X's method of offering the plan; and (b) Rate how strongly you would recommend Counselor X's argument. The name

of the counselor asked about is the same as the name in the vignette just presented; we used the term "Counselor X" only for the purpose of presenting the questions here. Two other open-ended questions were asked and scored: (a) What is your idea of a good counselor–patient relationship? and (b) Give the best reasons why that is a good relationship.

Procedure

Participants were contacted via a program listserv. In this e-mail, the study was described, and participants were provided with a link that they could use to directly link to the survey. In addition to participants' ratings on the counselor–patient vignettes, we also collected the following measures of student performance in the counseling program: A) Overall Student Grade Point Average; B) Student's GPA in selected "Hard" courses in the program (two courses, Research Methods and Differential Diagnosis are found by students to be particularly hard); C) Number of credits a student had completed, and D) Clinical Ratings: The Clinical Ratings were of a student's ethical skills and understanding by a Practicum/Internship supervisor. This was collected during their first clinical experience, called a Practicum. Using a 1 to 5 scale (1 = Far below Expectations to 5 = Far above Expectations) on such skills as knowledge of general ethical guidelines, knowledge of ethical guidelines of practicum placement, whether they demonstrated awareness and sensitivity to ethical issues, whether their personal behavior was consistent with ethical behavior, and whether they consulted with others about ethical issues if necessary. These were then averaged together to obtain an average rating in this area. At the time of this study, clinical ratings were available only for 7 of the students. This variable was dropped from the analysis to follow.

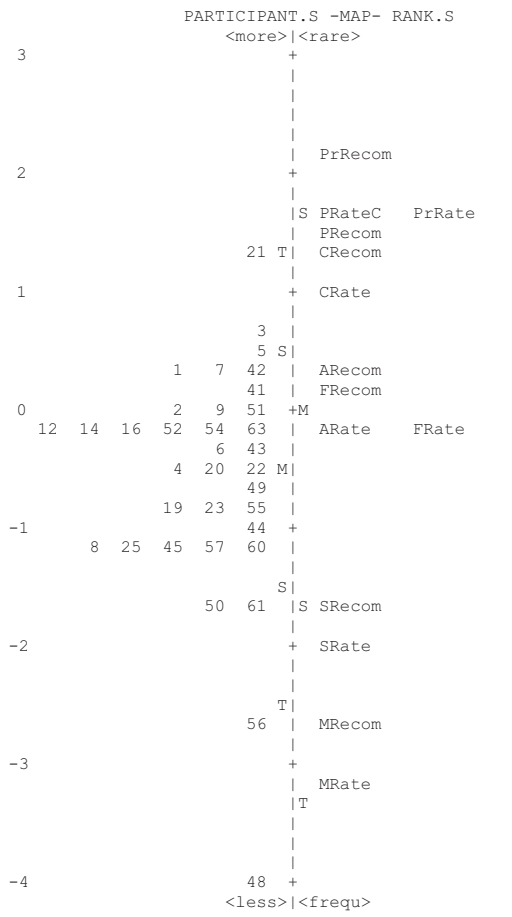

Figure 4. Rasch map of counseling patient vignette data.

Results—Study 2

The responses of 35 participants to the vignettes were first Rasch analyzed. The Rasch map is shown in Figure 4. As can be seen from the Rasch Map, the vignettes (also referred to as "items" in Rasch analyses) were scaled pretty much as predicted. Note that in Rasch Maps, items with more positive scores are those in which **less** of the characteristic being rated is found. In studies of the vignettes, which vary in hierarchical complexity, the dimension being rated is the hierarchical complexity. The Preoperational (Pr), Primary (P) and Concrete (C) Vignettes are at the top of the scale on the right hand side, showing that participants rated these counselors lower (or less hierarchically complex) than others and recommended them less. Furthermore, within these three orders, the Preoperational is generally lower than the Primary (except in one case in which they are at the same level), and the Primary is lower than the Concrete. The Abstract and Formal order vignettes are roughly in the middle of the scale with the Abstract vignettes being rated either slightly lower in the case of recommending the counselor, and at the same level as far as rating

the counselors. Finally, Systematic and Metasystematic vignettes are found toward the negative end of the scale, suggesting that these were rated higher (or more hierarchically complex) than counselors at the other orders and were also recommended more highly. The mean item score, shown by the M on the right side of the line down the middle was at the level where the Abstract and Formal items are clustered together. Based on its location, we assign the mean item stage to be 10.5 ($SD = 1.67$).

The participant scores, shown on the left side of the center line, ranged from 1.39 to −3.95, with $M = -.55$ and $SD = .97$. Although these scores are not converted into stage scores, it can be seen that the mean participant score of −.55 is transitional between Formal and Systematic. In addition, the largest number of participants ($N = 21$) are clustered between the Abstract/Formal orders and the Systematic Order. In this sample, and with this instrument, $N = 4$ (or 11.4%) of the participants were Systematic or above.

Finally, it was shown that the OHC of the items predicted the item scores (performance) with $r(35) = .949$ ($R^2 = .901$). That is, participants' ranking of the counselors in each vignette were scaled in a way that was a close match to the hierarchical complexity of the vignette, with the vignettes in which counselors behaved in a less hierarchically complex fashion being less preferred, and those in which counselors behaved in a more hierarchically complex fashion being more preferred. These results are shown mainly to establish that these data are similar to other counselor–patient data that has been collected (Commons et al., 2006).

Coding of Open-Ended Questions

The open-ended question "Give the best reasons why that is a good [counselor–patient] relationship" was coded using the MHC coding scheme. The mean stage was 11.38 ($SD = .74$; $Mdn = 11.5$), which places it in between Formal and Systematic. This mean is somewhat higher but not that different from the mean obtained on the vignettes. Scores had a more limited range on this instrument. Responses were coded from Stage 10 (Abstract), to State 12 (Systematic). We found 5 (15%) at the Abstract stage, 13 (39%) at the Formal stage, and 11 of the replies (33%) were coded at the Systematic stage.

The main interest of this study was to see whether the participants' performance in their counseling program was predicted by either Rasch Person Scores or the stage scoring of the Why Scores.

The intercorrelations of the variables are shown in Table 2. First, it should be noted that the two measures of stage—the Rasch Person Score, and the Why Score—were not significantly related to each other. When looking at the relationship of each of these stage measures to Overall GPA, only the Scoring of the Why question was significantly related to overall student GPA. Neither of these stage measures were related to students' grades in the two hard courses, nor were they related to the number of credits a student had taken.

A factor analysis was conducted on this same set of variables. These results are shown in Table 3. There were two factors found, accounting in all for 62.4% of the variance. The first factor (37.05% of the variance accounted for) appears to be the "standard academic success" factor, because both measures of GPA loaded highly on it. The stage scoring of the Why Question loaded positively as well. The second factor (25.36% of the variance) had the number of credits taken loading on it most highly. Rasch Person Score also loaded highly, although because higher stage scores on this are negative, and the loading is positive, this would suggest that the longer students spend in the counseling program (in terms of number of credits), the more likely they are to select a counselor with a lower stage method of informed consent.

Finally, we examined some regression models to see whether combining some of the vari-

Table 2
Correlation Matrix Showing the Bivariate Relationships Among Rasch Person Score, Scoring of the Why Question, GPA, and Credit Measures

Variables	1	2	3	4	5
1. Rasch Person Score	—				
2. Scoring of Why Question	−.024	—			
3. Overall GPA	−.114	.369*	—		
4. GPA hard courses	−.065	.180	.650**	—	
5. No. of credits	.175	−.090	.321	−.24	—

* $p < .05$. ** $p < 01$.

Table 3
Factor Analysis of Outcome Variables, Such as GPA, and the Predictors, Rasch Person Score, and Stage Scoring of the Why Question

Variable	Components	
	1	2
Overall GPA	.903	.322
GPA hard courses	.822	−.203
No. of credits	.008	.923
Rasch Person Score	−.199	.519
Stage Why Question	.569	−.050

ables in the models would allow us to predict how well students did in the program in terms of overall GPA. We ran models both with and without number of credits. These regression equations are presented next. Standardized coefficients are presented.

To see how well the Rasch Person Score (RPS) predicted GPA:

1. $\text{GPA} = -.114\ \text{RPS}$
   $R^2 = .013,$ and $F(1, 29) = .37,$ ns

As expected, and as already shown in the correlation table, there is no relationship of Rasch Person Score to a person's GPA.

2. $\text{GPA} = -.176\ \text{RPS} + .352\ \text{credits}$
   $R = .365,\ R^2 = .133,$ and $F(1, 24) = 1.68,$ ns

Including number of credits in the equation did not result in a significant prediction of GPA by these two variables.

To see how well the scoring of the Why question predicted GPA:

3. $\text{GPA} = .369\ \text{Why}$
   $R^2 = .136,\ F(1, 29) = 4.25,\ p = .049*$

4. $\text{GPA} = .401\ \text{Why} + .357\ \text{credits}$
   $R = .512,\ R^2 = .262,$ and
   $F(1, 24) = 3.912,\ p = .035*.$

This second set of regression equations shows that students' overall GPA was significantly predicted by the scoring of the Why question, and that when the number of credits taken was also included, this increased the amount of variance accounted for.

Discussion—Study 2

In this study, it was found that the average score of participants from the counseling program was between Formal and Systematic stage, being just slightly closer to Formal than to Systematic. This was true whether or not they were assessed with the structured vignette instrument, or with stage scoring of responses to a Why question. These findings are completely consistent with the earlier work of Eriksen and McAuliffe (2006). Although the mean reply to the two instruments were different, when we looked at numbers of individuals at each stage, we found that 33% of individuals studied were scored at the Systematic level in their reply to a brief question about why a certain way of seeking consent was more effective, however, in reply to the vignettes, 11.4% of the participants were Systematic. This contrasts to findings of previous studies by Commons and colleagues (e.g., Commons & Ross, 2008), showing that roughly 20% of unselected online samples score as Systematic stage. This does raise the question as to whether more of these graduate students should have been reasoning at the Systematic level.

The two assessments used—the structured vignettes and the narrative reply that was scored—were not only somewhat different in how they measured stage, but the correlational analysis showed that they were completely unrelated. This is unexpected in light of some previous findings of Commons and colleagues (Giri, Commons, & Harrigan, 2014), showing that scores from instruments in different subdomains of science and mathematics were highly intercorrelated. These measures have also been shown to have a reasonably high relationship to structured methods of assessing social or moral development. There are two possible reasons for the finding that the two methods used in this study seem to be unrelated. One is simply that the range of scores on both measures was quite restricted, with large numbers of scores being clustered between the Abstract and Formal stages. It is also the case that the somewhat higher number of Systematic responses on the stage scoring of the Why ques-

tion may reflect the fact that in a graduate program students are trained mainly in writing their views about counseling related topics. As part of that training, they have likely written previously on the topic of what makes someone a good counselor, and why, but they also would have received feedback on their responses, allowing them to further elaborate on those. On the other hand, they would have had little training or exposure to structured instruments in which they are asked to rate counselors.

This could also explain why responses to the Why question were a better predictor of the student's Overall GPA, as seen in the regression analyses.

Conclusions

There are several ways in which the results from these two studies may be relevant. First, this helps researchers to begin to understand the usefulness of measures such as these for evaluating graduate students. Although the outcomes examined in this study were limited to GPA, there is a suggestion that at least some measures of stage, particularly those that are a good match for the kinds of activities that students engage in during graduate study (e.g., writing), can be useful in predicting how well students do in their graduate study. Clearly, studies that examine a wider range of graduate students are needed in order to provide a much more complete picture of the relationship of higher stages of development to success in graduate study.

Second, this is one of just a few studies that has examined the extent to which stage scores based on the MHC may have real world applications. Because the results show that those whose writing contains more higher stage statements also appear to do better on graduate school tasks that predict good grades, this provides a small amount of predictive validity for the model. Clearly, a number of different kinds of validation studies are necessary, especially those conducted with different instruments and large enough sample sizes, in order to generate more of an argument in this regard. This study suggests that such exploration could be worthwhile.

References

Basseches, M., & Mascolo, M. F. (2010). *Psychotherapy as a developmental process*. New York, NY: Routledge.

Commons, M. L., & Jiang, T. R. (2014). Introducing a new stage for the model of hierarchical complexity: A new stage for reflex conditioning. *Behavioral Development Bulletin, 19,* 1–8. http://dx.doi.org/10.1037/h0100582

Commons, M. L., & Richards, F. A. (1984). A general model of stage theory. In M. L. Commons, F. A. Richards, & C. Armon (Eds.), *Beyond formal operations: Vol. 1. Late adolescent and adult cognitive development* (pp. 120–140). New York, NY: Praeger.

Commons, M. L., Rodrigues, J. A., Adams, K. M., Goodheart, E. A., Gutheil, T. G., & 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.

Commons, M. L., & Ross, S. N. (2008). What post-formal thought is, and why it matters. *World Futures, 64,* 321–329. http://dx.doi.org/10.1080/02604020802301139

Commons, M. L., Trudeau, E. J., Stein, S. A., Richards, F. A., & Krause, S. R. (1998). Hierarchical complexity of tasks shows the existence of developmental stages. *Developmental Review, 8,* 237–278. http://dx.doi.org/10.1006/drev.1998.0467

Crone-Todd, D. E., & Gonsalves, J. (2010, March). *The use of the model of hierarchical complexity in scoring senior-level undergraduate writing samples*. Poster session presentation at the 25rd Annual Adult Development Symposium: Society for Research in Adult Development, Philadelphia.

Eriksen, K., & McAuliffe, G. (2006). Constructive development and counselor competence. *Counselor Education and Supervision, 45,* 180–192. http://dx.doi.org/10.1002/j.1556-6978.2006.tb00141.x

Fischer, L. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review, 87,* 477–531. http://dx.doi.org/10.1037/0033-295X.87.6.477

Giri, S., Commons, M. L., & Harrigan, W. J. (2014). There is only one stage domain. *Behavioral Development Bulletin, 19,* 51–61. http://dx.doi.org/10.1037/h0101081

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures*. New York, NY: Basic Books. http://dx.doi.org/10.1037/10034-000

Kjellström, S., Ross, S. N., & Fridlund, B. (2010). Research ethics in dissertations: Ethical issues and complexity of reasoning. *Journal of Medical Eth-*

ics, 36, 425–430. http://dx.doi.org/10.1136/jme.2009.034561

Leite, S. F. (2016). Successes in cultural evolution raises the variability in humans' highest stage attained. *Behavioral Development Bulletin, 21,* 165–175. http://dx.doi.org/10.1037/bdb0000033

Li, E. Y., Commons, M. L., Miller, J. G., Robbinet, T. L., Marchand, H., Ost, C. M., & Ross, S. N. (2014). Relationship among measures within the social and moral developmental domain. *Behavioral Development Bulletin, 19,* 106–111. http://dx.doi.org/10.1037/h0100595

Mingee, C. M. (2013). Retention of a Brightness Discrimination Task in Paramecia, (P. caudatum). *International Journal of Comparative Psychology, 26,* 202–212.

Pascual-Leone, J. (1970). A mathematical model for the transition rule in Piaget's developmental stages. *Acta Psychologica, 32,* 301–345. http://dx.doi.org/10.1016/0001-6918(70)90108-3

Rest, J., Narvaez, D., Bebeau, M. J., & Thoma, S. J. (1999). *Postconventional moral thinking: A neo-Kohlbergian approach*. Mahwah, NJ: Erlbaum.

Ross, S. N., Commons, M. L., Li, E. Y., Stalne, K., & Barker, C. D. (2014). Toward defining order 16 and describing its performance for the model of hierarchical complexity. *Behavioral Development Bulletin, 19,* 33–36. http://dx.doi.org/10.1037/h0100588

Received June 14, 2016
Revision received September 19, 2016
Accepted September 19, 2016 ■