Running head: MEASURING COMPLEXITY OF LEARNING TASKS

Development of an Instrument for Measuring the Complexity of Learning Tasks

Rob J. Nadolski, Paul A. Kirschner, Jeroen J. G. van Merriënboer, and Jürgen Wöretshofer

Open University of the Netherlands

Date of first submission, October 23, 2002

Date of resubmission, March 27, 2003

Date of final version, September 23, 2003

Author Notes

The authors thank the law teachers and students from various Dutch universities for their participation in this study and especially their colleagues Martin Baks and Dick H. van Ekelenburg from the faculty of Law at the Open University of the Netherlands who did a complex task in the development team.

Correspondence concerning this article should be addressed to Rob J. Nadolski, Educational Technology Expertise Center, Open University of the Netherlands, Valkenburgerweg 177, 6419 AT Heerlen, The Netherlands. Electronic mail: rob.nadolski@ou.nl

Abstract

An instrument for measuring the complexity of learning tasks in the field of Law was developed and tested in three experiments. In Experiments 1 and 2, teachers used the card-sort method to rate the complexity of learning tasks. Based on the outcomes, a benchmark scale with four criterion tasks was used in Experiment 3. The results showed the benchmark instrument to be valid and easy to use, allowing instructional designers to design competency-based learning environments that better take task complexity into account. The Instructional Design model, in which teachers determine task complexity, is briefly described.

Development of an Instrument for Measuring the Complexity of Learning Tasks

Competency-based Multimedia Practicals (CMPs) provide realistic situations in which meaningful learning takes place in a self-contained electronic learning environment. CMPs provide context relevant practice to students and belong to the collection of so-called competency-based learning environments used for attaining complex skills (Brown, Collins, & Duguid, 1989; Nadolski, Kirschner, van Merriënboer, & Hummel, 2001; Westera & Sloep, 1998). Competencies always include complex skills because different constituent skills need to be performed in an integrated and coordinated fashion in order to deal with realistic situations. The realistic situations or learning tasks in competency-based learning environments are often too complex for learners to deal with, without some form of simplification. To alleviate this complexity problem the learning tasks are typically non-systematically decomposed into simpler tasks that are within reach of the learners' capabilities. An underestimation of task complexity results in too little decomposition and too complex tasks for learners to deal with, hampering the attainment of the integrated set of constituent skills. An overestimation results in too much decomposition and in too easy and therefore non-challenging tasks (Bonner, 1994). In addition, an overestimation of task complexity results in the development of unnecessary instructional materials resulting in higher development costs. The central question is then: How can we determine task complexity so that we can achieve optimal decomposition of learning tasks?

Task complexity has both an objective and a subjective component. Objective task complexity results from the characteristics or the nature of the task itself. Subjective complexity is determined by the characteristics of the task and of the person carrying out the task. Playing an etude from Chopin, for example, is objectively more complex than practicing the scales on a piano. This is 'objectively' true for both the expert and the novice, although the expert will

'subjectively' experience playing Chopin as being less complex than the novice will. In a more cognitive vein, sentence complexity is another example of where complexity can be objectively determined, irrespective of the readers' familiarity with the content of the sentence. Most often, sentence complexity and readability are determined on the basis of sentence length, word length, number of phrases and clauses, et cetera (e.g., Flesch, 2003; Vaso, 2000). Although such readability formulas are not undisputed (Brandle, 2002; Clough, 2000; Pikulski, 2002), one would agree that a sentence in which the subject and object are separated by a large number of dependent and independent clauses and where the average word length is quite long is more complex and thus more difficult to understand than a simple sentence. Again, the experienced reader will have an easier time than the novice (subjective), but this does not nullify the fact that sentences also differ objectively.

Tasks that consist of higher-level unique constituent skills requiring more coordination have higher objective complexity than tasks with fewer unique constituent skills requiring less coordination. Subjective task complexity is the complexity experienced by the learners while performing the task as a reaction to the task characteristics, their own characteristics and the characteristics of the environment.

Studies have shown that task complexity can be used to predict task performance. This is true for both objective task complexity (e.g., Boggs & Simon, 1968; Early, 1985; Kernan, Bruning, & Miller-Guhde, 1994; Scott, Fahr, & Podsakoff, 1988) and subjective task complexity (e.g., Huber, 1985; Taylor, 1981). While these studies focused on either objective or subjective task complexity, a recent study by Maynard and Hakel (1997) explicitly focused on uncovering the relationships between the two. What they found was that objective task complexity is a good predictor of subjective task complexity, in the sense that higher levels of objective task

complexity lead to higher levels of subjective task complexity. In addition, their research showed a high correlation between perceived (subjective) and objective task complexity, a finding consistent with results from earlier studies (Huber, 1985; Kernan et al., 1994; Scott et al., 1988).

The present study concerns the development of a reliable, valid and easy to use measurement instrument for rating the objective complexity of Law learning tasks. Several domain-independent instruments have been developed to determine objective task complexity (e.g., Bonner, 1994; Byström & Järvelin, 1995; Campbell, 1988; Campbell & Gingrich, 1986; Wood, 1986). The main problem with these instruments is that they are difficult to use and usually involve considerable training. Wood (1986), for example, has developed an instrument that makes total task complexity operational by distinguishing between component complexity, coordinative complexity, and dynamic complexity of a task. Component complexity is a direct function of the number of distinct acts executed in the performance of the task and the number of distinct information cues processed in the performance of those acts. Coordinative complexity refers to the nature of the relationships between task inputs and task products, the nature of the relationship is given between 'n' task input(s) and 'm' task output(s) (n =1, 2…..; m = 1, 2……). Dynamic complexity refers to how often individuals must adapt to changes in the cause-effect chain or in the means-ends hierarchy for a task during the performance of a task, due to changes in the world which have an effect on the relationship between task inputs and products. For instance, a pilot when landing a plane has to respond to changing weather conditions, the height above sea level and the height above the landing strip. Application of Wood's model requires determination of these three types of complexity and weighting factors for each of them in order to finally determine task complexity.

Campbell (1988) has offered an approach that has shown to have empirical value in the field of business administration curricula. According to him, task complexity is directly related to those task characteristics that increase information load (i.e., the number of dimensions of information requiring attention), information diversity (i.e., the number of alternatives associated with each dimension), and/or the rate of information change (i.e., the degree of uncertainty involved). He identifies four basic dichotomous task characteristics that affect load, diversity and/or change namely the presence or absence of: (1) multiple potential ways ("paths") to arrive at a desired end-state; (2) multiple desired outcomes to be attained; (3) conflicting interdependence among paths to multiple outcomes, and (4) uncertain or probabilistic links among paths and outcomes. On the basis of these four characteristics, sixteen task-types can be distinguished (presence/absence of each of the four task characteristics). However, an exact ordering of tasks from simple to complex is difficult because Campbell does not specify the relative contribution or weight of each of the four basic attributes.

Burtch, Lipscomb, and Wissman (1982) described a simpler benchmark scaling technique in which anchor tasks are used to describe each complexity level on a scale. New tasks are compared to the anchor tasks and the best likeness determines the complexity. This is similar to the Mohr-scale for determining the hardness of minerals where a mineral is scratched with the ‘anchor’ minerals for comparison; the harder mineral leaves a scratch on the softer one. Such an instrument requires very little learning and training; subject matter experts in the task domain can easily use the instrument if the expected prior knowledge of task performers or learners has been defined. Our research has applied this general approach for the development of our benchmark instrument using a conceptual frame of reference largely based on Merrill's Component Design Theory (1987).

As was seen from the description of earlier approaches (Wood, 1986; Campbell, 1988) determining weighting factors for the relative contribution of the various attributes to objective complexity is often difficult or even unknown (Campbell, 1988). Since our conceptual frame of reference centers on learning tasks and concerns intellectual operations involved in learning we attempted to alleviate this problem by using Merrill's Component Design Theory (1987) which distinguishes four major hierarchical categories of operations ("performances") that can be defined as the four levels of complexity (i.e., very simple, simple, complex, and very complex). All levels are relative to the prior knowledge of the learners because they are based on the unfamiliarity of the learner with the learning task in which this operation occurs. Once the learner has mastered the learning task in question, this same task becomes a routine task, therefore becoming simpler than it was before. Complexity increases from (1) remember an instance: gain and remember facts / retention (very simple); (2) understand a generality: gain generalized, abstract knowledge / insight or understanding (simple); (3) use: apply knowledge in familiar settings (complex); and (4) find: apply knowledge in unfamiliar settings / problem solving and qualitative reasoning (very complex). The intellectual operations in our frame of reference are hierarchically ordered with each higher level subsuming the previous ones. But as is the case for the discriminating characteristics in Campbell's model, an exact ordering of complexity remains difficult since the relative contribution of each of the four classes of intellectual operations to a particular learning task is unknown while the breadth of a certain class of intellectual operations can be very large. This means that under certain circumstances understanding a generality (e.g., understanding the concept Justice) can be more complex than using knowledge (e.g., applying a simple procedure for determining the maximum punishment for a certain crime).

This conceptual frame of reference was used for the development of a benchmark instrument for measuring the complexity of Law learning tasks. The development entailed carrying out three related experiments. The next section describes the general methodology of all three experiments.

General methodology

A similar methodology was used for all three experiments. Where relevant, differences are given when separate experiments are described.

Participants

All participants were informed about the experiments, the time schedule and the estimated workload and were compensated with a small gift plus a small monetary remuneration (circa $80) per experiment. Two groups of participants were used. One group was composed of Law teachers at Dutch universities from the fields of Criminal Law and Civil Law (n = 33). The second group was composed of graduate level Law students at Dutch universities (n = 12).

For the first group, 23 teachers working at different Dutch universities registered before the start of Experiment 1. Ten additional teachers registered while conducting Experiment 1. No participants from the teacher-group participated in both Experiments 2 and 3 since Experiment 3 included tasks from Experiment 2. Teachers participated in two experiments maximally (Experiments 1 and 2 or Experiments 1 and 3). Participants from the student-group (n = 12) only took part in Experiment 3.

Material development

The basis material used in this research was taken from existing Law courses, some of which were competency-based.

The--to be rated--Law learning tasks for the various experiments were restricted using two simple guidelines. First, the tasks were suitable for sophomore Law students. Since all Dutch universities have almost identical Law curricula for the freshman year, all sophomore students can be expected to have comparable prior knowledge and thus the Law teachers could be expected to have similar views of what these students should be able to do. Tasks from exotic sub domains of Law were also excluded. Second, the length of the tasks (formulation plus solution in keywords) was standardized so that "length of task" would not be a contaminating artifact in the determination of complexity.

The two guidelines in conjunction with the conceptual frame of reference were used to determine 56 tasks to be included in the various experiments. All four members of the development team (two criminal law teachers, one civil law teacher and one educational technologist specialized in Law courses) independently scored the complexity of the tasks on a 4-point ranking scale (very simple, simple, complex, very complex). The conceptual frame of reference which formed the criteria for determining the complexity was known to all of them. There was no simple algorithm for applying the frame of reference so it was possible for the developers to apply the criteria differently. For all 56 tasks, Cohen's Kappa was calculated (K = 1 for 46 tasks, K = .7 for 10 tasks). After rating, the development team discussed their ratings for further articulation of their conceptual frame of reference.

Procedure

All printed materials (including instructions) were sent to the participants' work addresses. They had ten workdays to return the materials in a stamped self-addressed envelope. They were informed that they should work individually and that it would take them approximately three hours to do the necessary work. Participants were offered the opportunity to

receive further information (by mail or phone). In all three experiments no participant made use of this offer. A reminder was sent when the deadline for return had expired. Upon the completion of an experiment, participants were thanked for their participation, received their gifts and were informed about their participation in the upcoming experiments.

Experiment 1

One important criterion for an easily usable benchmark scale is its non-specificity for raters' area of expertise. Experiment 1 studied whether the specific expertise of a participant in a sub domain of Law (Civil or Criminal) influenced the rating of tasks from their own or from the other sub domain. The experiment was also used to begin the process of determining anchor-tasks for further experiments and as a pilot for the design of the questionnaire to be used in the further experiments.

Method

Participants

Nineteen law teachers (7 Criminal Law, 12 Civil Law) employed at Dutch universities (10 distance education, 9 face-to-face education), returned their results (response rate = 83%).

Materials

The materials consisted of Criminal Law and Civil Law learning tasks in two separate packages plus a series of questionnaires. Each task package contained 16 Law learning tasks selected from the original 56, one task per page. The instrument for gathering the data consisted of seven different parts:

1. Card sort task for complexity. Participants were asked to sort each of the 16 tasks into four equal piles with comparable complexity (very simple, simple, complex, very

complex). The tasks provided had--according to the development team--an equal distribution within the conceptual frame of reference (i.e., four tasks for each category).

2. Task ranking within piles. Once participants had made the four equal piles, they ranked the tasks within each pile from least to most complex. As a result, for both sub domains, the 16 tasks were sorted with respect to increasing complexity on a 16-point ranking scale.
3. Students' time on task estimations. Participants indicated how long they felt it would take a sophomore to learn to perform each task: this 'learn to perform' is stressed as for instance 'to learn to perform a plea' is more time-consuming than 'to perform a plea'. Time to conduct a task is considered by some researchers to be a good indicator of task complexity (Maynard & Hakel, 1997; Winne, 1997).
4. Rating criteria. To determine their conceptual frame of reference for determining task complexity, participants were asked to rate 18 assertions on possible criteria for judging the complexity of Law learning tasks on a 4-point categorical scale ranging from totally disagree (1) to totally agree (4). Assertions dealt with topics such as 'amount of possible solutions', 'kind of intellectual operations required', et cetera. There was space left for the participants to add other topics.
5. Participants' time on task. To determine the speed of use of the different instruments, the time needed to carry out the 'card-sort and ranking'-task as well as for 'estimating the students' time on task'-task was reported by participants.
6. Ease of use. Since speed of use is not necessarily the same as ease of use, a 9-point categorical scale developed by Paas and van Merriënboer (1994) was used to measure

the perceived cognitive load of the (1) card-sorting task, (2) the ranking task, and (3) the 'estimated students' time on task'-task. Cognitive load is supposed to be an indication for ease of use; the less mentally demanding the task, the lower the cognitive load. This was included to check the perceived cognitive load of what the participants were asked to do and thus to check if the instrument is easy to use.

7. General information. Data were collected on participants' experience, gender, et cetera.

Design and procedure

A 2x2 (expertise x sub domain) completely crossed, factorial design was employed. The expertise of the rater could be in Criminal Law or Civil Law as could be the sub domain of the learning tasks.

Participants were asked to sort the learning tasks provided (formulation plus solution in keywords) with respect to their judgment of the complexity for sophomore Law students to learn to carry them out. It must be stressed here that the respondents did not rate how complex it would be to carry out the task, but rather how complex it is to LEARN how to carry out the task. For example, learning to walk a tightrope is a complex task, whereas once having mastered this, it becomes quite easy for the tighrope-walker. Tasks were randomly ordered for the card sort. Task ranking within the four piles and students' time-on-task estimation followed this.

Results

Participants' expertise

We expect that the specific field of expertise of participants would not influence their ratings for the sub domains, since all participants had experience with all offered tasks during their own study. In other words, a teacher of Criminal Law would also be familiar enough with

sophomore Civil Law learning tasks to rate them with a result similar to the teacher of Civil Law teacher. Secondly, since all freshman law curricula are (almost) identical at all Dutch universities and all faculty members at the Open University of the Netherlands (distance education) are products of "traditional" face-to-face universities, we expect that the "type of university" of the participant (distance education vs. face-to-face) also would not influence their ratings.

A univariate analysis of variance for the sum of deviations of participants' ratings to the conceptual frame of reference revealed no significant differences in participants' ratings for Criminal Law versus Civil Law learning tasks based upon their area of expertise (Criminal Law tasks, F (11, 6) = .029, MS = .303, p = .867; Civil Law tasks, F (11, 6) = .004, MS = .063, p = .948). For all tasks taken together the area of the participants' expertise did not influence their ratings for Criminal Law learning tasks and Civil Law learning tasks. The results for all separate tasks also showed the same pattern. Participants also indicated that they did not expect themselves to rate tasks in their own sub domain of expertise better, as confirmed by the rating results (cf. self-efficacy: Bandura, 1982). Participants regarded their 'teaching expertise' of slightly more importance than their 'subject matter expertise' for the quality of their ratings. This difference, however, was not significant.

A second univariate analysis of variance for the sum of deviations of participants' ratings to the conceptual frame of reference showed that "type of university" also did not affect ratings for both groups of tasks as a whole (Criminal Law tasks, F (12, 6) = .067, MS = .936, p = .799; Civil Law tasks, F (12, 6) = .038, MS = .395, p = .848) nor for all separate tasks.

Card sort

To estimate the extent to which the individual ratings of the participants in the card-sort tasks correspond with each other, the concordance coefficient--Kendall's W--was calculated

(Hays, 1981; Siegel, 1956). This coefficient was calculated for all four conditions and for both the 16-points ranking scale and 4-points ranking scale (Table 1).

****

Insert Table 1 about here

****

All coefficients are significant at the 1% level of probability confirming that the participants showed a large degree of agreement on the rankings and ratings and that the participants were applying the same standard in ranking the tasks under study.

****

Insert Table 2 about here

****

Table 2 presents the descriptive statistics for the separate tasks in the card sort and estimated student's time on task; the latter can be disregarded for the moment. The order of the tasks in the card sort from very simple to very complex was determined by the mean rating scores. The classification of a task in one of the four categories on the basis of the mean score or on the basis of the median is the same for all tasks. The data showed that participants' ratings for the separate tasks differed quite a lot. If differences occurred between raters' classifications and the conceptual frame of reference, the deviation was maximally one class. The data presented in Table 2 show that the consensus among participants for the extremes (very simple tasks and very complex tasks) was larger as for the two intermediate categories. The rating-values based on the median (Mdn) and the values based on the conceptual frame of reference (Rf) were much more in correspondence for both very simple tasks and very complex tasks than for the other two categories. The correspondence with the conceptual frame of reference was 87.5% for very

simple tasks, 62.5% for very complex tasks, but only 37.5% for simple tasks and 25% for complex tasks. This pattern was observed for both the Criminal Law and the Civil Law tasks. The tasks, based on their confidence scores (P(c=ci)), could not be clearly attributed to one complexity class; only Criminal Law task 3 (cr3) could be attributed with high confidence ($p < .1$) in one category, namely very complex. Here too there was more consensus for tasks on the extremes of the scale (belonging to either very simple or very complex) than for the two middle categories (either simple or complex).

Estimated students' time on task

Descriptive statistics on students' time on task show participants' ratings for the separate tasks again differing greatly (Table 3). Spearman's correlation between the ranks from the card sort and the ranks from estimated students' time on task was .785 ($p < .01$) for the Criminal Law tasks and .921 ($p < .01$) for the Civil Law tasks (Table 3). Thus, the complexity rankings resulting from estimated students' time on task and card sort were very similar.

****

Insert Table 3 about here

****

Rating criteria

From the results of participants' scores on the 18 assertions about criteria they used for rating the complexity of the learning tasks their--collective--conceptual frame of reference for rating could be derived. Means for those criteria on the 4-point scale ranged from 2.21 (SD = 0.86) to 3.68 (SD = 0.48).

From these data, the three most important criteria for their ratings were: (a) quantity of information searched for and combined (M = 3.68, SD = 0.48), (b) quantity of information given

and combined (M = 3.47, SD = 0.52), and (c) kind of intellectual operations required (M = 3.42, SD = 0.61).

Participants' time on task

It took the raters approximately five minutes to evaluate each task, including the time needed to read the task (for 16 tasks: M = 73.3 min, SD = 12.5 min). The time needed to estimate 'students' time on task' was about one minute for each task (for 16 tasks: M = 16.7 min, SD = 2.6 min). Here the task was conducted after the ranking-task so the reading time of the task was not taken into account.

Ease of use

Cognitive load on the 9-point categorical scale (1 = very easy, 5 = not easy, not difficult, 9 = very difficult) can be used as an indication for the ease of using the instruments. Perceived cognitive load values were collected for the card-sorting task (M = 5.68, SD = 1.77), the ranking task (M = 4.47, SD = 1.65) and the 'estimated students' time on task'-task (M = 6.32, SD = 1.89). Cognitive load values showed that students' time on task estimations cause the highest load, but comparing the mean cognitive load values in an independent samples t-test showed that this task was not significantly more mentally demanding than the other tasks (card-sort and ranking). All tasks were low to moderately mentally demanding for participants.

Discussion

Results show that neither specific expertise nor type of university of the raters influences their ratings. Based upon these findings, we concluded that we could use experts from both sub domains, from different types of universities, and tasks from both sub domains in the following experiments.

Since the participants did not regard their teaching expertise to be significantly more important than their subject matter expertise, one might ask whether raters necessarily be teachers. If, for example, graduate level Law students make similar ratings to Law teachers, this would allow raters to be more easily recruited and would make the rating process less costly. The effects of teaching expertise versus subject matter expertise were studied in Experiment 3.

Participants' rankings for estimated students' time on task showed results similar to their rankings from the card sort, a result that is consistent with the findings by other researchers (Maynard & Hakel, 1997; Winne, 1997). In our view, time-estimations can only be used to determine relative task complexity and not absolute task complexity since tasks in the field of Law often require reading large amounts of information. In other words, time-on-task estimations can only be used to predict the complexity of a learning task provided one avoids a rating artifact as 'length of task'. This criterion was met in our first experiment.

The task-complexity ratings showed too much variance for confident rating. Separate tasks in the intermediate categories had the lowest confidence values; the extremes showed more consensus. Participants may need more explanation on these categories for more uniform ratings. Providing anchor tasks might improve the consensus in their ratings of the separate tasks. Experiment 2 was designed to determine such anchor tasks.

The highly significant value of Kendall's coefficient (W) in the card-sort task shows that the participants sorted the items on the basis of the same criteria. Based upon the results obtained on rating criteria, the conceptual frame of reference used by the participants appears to coincide with our conceptual frame of reference, which primarily centers on the criterion 'kind of intellectual operations required'. Participants indicated this among their three most important

criteria for their ratings. Their ratings were quite often completely in line with our conceptual frame of reference and never showed a deviation of more than one class.

The results on time needed for rating and ease of instrument use show that they were able to make their ratings quickly and do not find this task to be mentally demanding. This is very encouraging since our goal is to develop an instrument that is quick and easy to use.

Experiment 2

Experiment 2 was carried out to determine anchor tasks that could be used in the upcoming experiment. Such anchor tasks to benchmark a complexity category are expected to positively support the rating process. Since the second experiment was conducted in a fashion largely similar to the first, the methodology will be treated in less detail.

Method

Participants

Twelve Law teachers (8 Criminal Law, 4 Civil Law) employed at Dutch universities returned their results in this experiment (response rate 80%). Nine of them also participated in Experiment 1.

Materials

The materials now consisted of descriptions of 24 Law learning-tasks (20 Criminal Law, 4 Civil Law; none of which were used in the previous experiment) and a series of questionnaires similar to Experiment 1. The tasks provided had--according to the development team--an equal distribution within the conceptual frame of reference (ie., six tasks for each category). The questionnaire for gathering data on the participants' conceptual frame of reference for rating was adapted for this experiment. Nineteen assertions (one new) were now scored on a 6-point

categorical scale to allow for more sensitive analyses. In this second experiment participants had to sort 24 tasks instead of 16.

Design and procedure

The design and procedure was the same as for the first experiment.

Results

Card sort and anchor tasks

Kendall's W was calculated for both the 24-point ranking scale ($W = .821$, $p < .01$) and the 4-point ranking scale ($W = .646$, $p < .01$).

The rating results for the separate law learning tasks to choose anchor tasks are presented in Table 4. The second--right--half of this Table presents results from Experiment 3 and can be disregarded at this moment.

****

Insert Table 4 about here

****

The order of the tasks, from very simple to very complex, was based upon the mean rating scores. The classification of a task in one of the four categories on the basis of the mean score or on the basis of the median were, except for task 24, the same for all tasks. The data showed that the ratings for the separate tasks differed quite a lot making it impossible to select more than one anchor task per intermediate category.

Tasks 10 and 1 could be attributed with high confidence ($p > .95$) to category 1 (very simple). Task 10 was chosen as anchor task for this category since task 1 might be too obvious as representative for this category. Both tasks 16 and 2 could be attributed with high confidence to category 4 (very complex) ($p > .95$). Task 2 was chosen as anchor task for this category. For the

intermediate categories, the task with both the highest probability of correctly belonging to a category (P(c=ci)) and where the rank based upon the mean and the conceptual frame of reference (Rm=Rf), was used as anchor task. For category 2 this was task 5 and for category 3 task 17. It was not possible to choose anchor tasks for those categories with a confidence level of 90% or higher.

Estimated students' time on task

Spearman's correlation between the ranks from the card sort and the ranks from estimated students' time on task was .95 ($p < .01$). Again, the complexity rankings resulting from estimated students' time on task and the card sort were very similar.

Rating criteria

The means on the 6-point scale of the participants' scores on 19 assertions about criteria for rating the complexity of the learning tasks ranged from 3.08 (SD = 1.31) to 5.33 (SD = .65). The three most important criteria for their ratings were: (a) kind of intellectual operations required (M = 5.33, SD = 0.65), (b) quantity of information searched for and combined (M = 5.08, SD = 0.67), and (c) quantity of juridical judgment (M = 4.92, SD = 0.79).

Participants' time on task

The 'card sort and ranking'-task took approximately four minutes per Law task, reading time included (for 24 tasks: M = 104.1 min, SD = 51.4 min). Estimating 'students' time on task' took less than one minute for each Law task (for 24 tasks: M = 15.5 min, SD = 7.2 min). As was the case in Experiment 1, this task was conducted after the ranking-task so the reading time was not taken into account.

Ease of use

Perceived cognitive load values for the card-sorting task (M = 6.08, SD = 1.31), the ranking task (M = 5.67, SD = 1.67) and the 'estimated students' time on task'-task (M = 6.00, SD = 1.65) were collected. Comparing those mean cognitive load values in an independent samples t-test showed that all tasks were of comparable load and imposed a moderate mental demand on the participants.

Discussion

The most important result was that all of the findings from Experiment 1 were replicated in Experiment 2.

Despite the highly significant value of Kendall's W, the results in this experiment again show that participants differed considerably in their ratings for the separate tasks, especially for the tasks belonging to the intermediate categories. As shown in Experiment 1, these differences cannot be attributed to participants' expertise.

The criteria participants indicated as most important --as was the case in Experiment 1-- closely resemble those that the development team used for constructing the conceptual frame of reference. Nevertheless, the differences of participants' scores within this conceptual frame of reference become especially manifest for categories 2 and 3. This is most probably the result of the earlier discussed breadth of intellectual operations within these two classes, which makes it difficult to unequivocally attribute a task to a certain complexity. Another complication is that the relative contribution or weight of each of the four discriminated intellectual operation classes to complexity is unknown. Using anchor tasks to benchmark a category could help the classification.

Experiment 2 was needed to choose anchor tasks for each category. For categories 1 and 4 this choice could be made confidently. For the two intermediate categories (2 and 3) less representative anchor tasks had to be chosen.

Experiment 3

Experiment 3 had three purposes. The first purpose was to determine the effect of providing anchor tasks for rating Law tasks. To this end, the results of the teacher group of Experiment 2 (without anchor tasks) were compared with the teacher group of this experiment (with anchor tasks). It was expected that using anchor tasks would result in more agreement about the ratings of the separate tasks. The experiment was also used to determine whether additional (better) anchor tasks could be identified. Finally, it investigated whether Law teachers and graduate level Law students would rate sophomore Law learning tasks in a similar manner. It was hypothesized that they would make similar ratings since all participants had encountered the to-be-rated tasks during their own study, and that the lack of experience of the graduate students would be compensated by the fact that they only recently covered the material.

Method

Participants

Two groups of teachers (Group 1, 2) and one group of graduate students (Group 3) were involved in Experiment 3. The 12 law teachers (8 Criminal Law, 4 Civil Law) from Experiment 2 constituted Group 1 and functioned as a control group here. Group 2 consisted of an additional 13 law teachers (7 Criminal Law, 6 Civil Law) employed at Dutch universities, six of them also participated in Experiment 1. Group 3 consisted of 12 graduate level Law students who had not

yet chosen a specialty. Groups 2 and 3 were used to measure the effect of the anchor tasks on the rating results, with Group 1 functioning as control group (no anchor tasks).

Materials

As four anchor tasks were chosen out of 24 tasks in Experiment 2, the remaining 20 tasks were used. Groups 2 and 3 received materials identical to the previous experiment with the exception of a separate package with four anchor tasks clearly labeled with the category of complexity they represented. The participants were also questioned on the perceived usefulness of the anchor tasks for rating.

Design and procedure

The procedure was, with exception of the anchor tasks, the same as for the preceding experiments except that the participants from Groups 2 and 3 were instructed to use the anchor tasks when carrying out the 'card-sort' and 'task-ranking' tasks.

Results

The results for participants' time on tasks and ease of use concurred with the earlier experiments and will not be separately reported here.

Anchor tasks

The card sort rankings constructed on the basis of the mean ranking scores for the separate tasks for both groups of teachers (Group 1, 2) were compared in a non-parametric correlation test (Spearman's correlation = .919, $p < .01$). Both groups of teachers made similar card sort rankings for the learning tasks included in both experiments. The anchor tasks (right half of Table 4) affected neither the variance in rating results nor the confidence in the ranking.

Card sort

Kendall's W was calculated for both the 20-point ranking scale (W = .693 (teachers), W = .791 (students)) and the 4-point ranking scale (W = .628 (teachers), W = .671 (students)). All coefficients are significant at the 1% level of probability.

Table 4 shows the rating results for the separate tasks by both groups of teachers. The left half presents the scores of the teachers in Experiment 2; the results for the teachers and students in Experiment 3 are presented in the right half. The consensus among teacher-participants for very simple and very complex tasks was--as was the case in the previous experiments--larger than for the two intermediate categories. The consensus was not improved by the anchor tasks. It was still not possible to choose anchor tasks for the intermediate categories with a confidence level of 90% or higher. The correspondence with the conceptual frame of reference was 80% for tasks in category 1, 100% for tasks in category 4, but only 40% for tasks in category 2 and 60% for tasks in category 3.

Participants' experience

Spearman's correlation coefficient between the ranks from the card sort for Law teachers and graduate level Law students was .968 ($p < .01$), showing that both groups rated the tasks very similarly.

Estimated students' time on task

Spearman's correlation between the ranks from the card sort and the ranks from estimated students' time on task was .904 ($p < .01$) for teachers (Group 2) and .961 ($p < .01$) for students. Thus, complexity rankings resulting from estimated students' time on task and the card sort were for both groups very similar.

Rating criteria

The means for the teachers' (Group 2) on the 19 assertions on rating criteria ranged from 2.85 (SD = 1.07) to 4.92 (SD = 0.64). The students' means ranged from 3.00 (SD = 0.85) to 5.17 (SD = 0.83).

The three most important criteria for the teachers' were: (a) kind of intellectual operations required (M = 4.92, SD = 0.64), (b) sub domain of law (M = 4.92, SD = 0.86), and (c) quantity of information searched for and combined (M = 4.92, SD = 1.04). Students used the criteria: (a) quantity of information searched for and combined (M = 5.17, SD = 0.83), (b) kind of intellectual operations required (M = 5.08, SD = 0.67), and (c) own experience as a student with these tasks (M = 5.00, SD = 0.95). Both groups emphasize the criterion around which the conceptual frame of reference is primarily centered (i.e., intellectual operations). There were no significant differences in their scores on 18 of the 19 assertions. A t-test for independent samples showed a significant difference between the groups for the means of 'own experience as a student with these tasks' ($t(23) = -3{,}18$, $p < .01$); graduate level students rated this criterion significantly higher than teachers.

Participants' opinion on the usefulness of anchor tasks

Statistical characteristics for participants' scores on the 6-point scale on the usefulness of the anchor tasks for their ratings are presented in Table 5. Participants regarded the anchor tasks to be of only limited use.

****

Insert Table 5 about here

****

Discussion

Experiment 3 investigated whether the use of anchor tasks would result in more confident ratings for the individual tasks. This appeared not to be the case for the teachers. It is not certain whether this also holds for graduate level students, as we did not collect students' rating results without the use of anchor tasks. Again, participants differed in their ratings for the separate tasks, especially for the tasks belonging to the intermediate categories.

Experiment 3 also did not result in discovering additional or better anchor tasks. As it was not feasible to unequivocally identify representative anchor tasks for the intermediate categories in Experiment 2, one could not expect a panacea from using them in Experiment 3.

Raters thought all anchor tasks to be of not much help in classifying the learning tasks. This is supported by the data that show that the classifications of the tasks for all categories did not have less variance when anchor tasks were available.

A positive result is the finding that graduate level students can apparently be used for rating learning tasks. Students rank the tasks in a similar way and report using a similar conceptual frame of reference for rating, which was also in line with the conceptual frame of reference as was used by the researchers. This simplifies the work of the educational developer for determining the complexity of learning tasks in CMPs since the "pool" of students is larger than the "pool" of teachers and the fees that need to be paid are lower. Teaching experience does not result in different or more valid ratings. The most plausible explanation for this finding is that graduate students' more recent experience with the tasks and content compensates for missing teaching experience and/or depth of knowledge. This is also in line with the finding that students' report that their own experience with these tasks is one of the three most important rating criteria. Teachers report this criterion to be significantly less important for their ratings.

The results of this experiment reconfirm the results from earlier experiments on card sort rankings and estimated student-time on task rankings.

General discussion

The most important conclusion of this study is that the instrument developed is reliable, easy to use, and does not require any specific training. Results from all three experiments show that raters apply the same standard in ranking learning tasks, using criteria similar to those in our conceptual frame of reference. The most relevant expertise for rating the learning tasks is the raters' own experience as a student with the tasks. Ratings were not influenced by the area of expertise or type of university (Experiment 1). Graduate level Law students and Law teachers also rated the tasks similarly (Experiment 3). Complexity rankings based on estimated students' time-on-task and card sorting are similar. Both methods of ranking are relatively fast and easy to carry out. Provided that length of task is controlled for estimating students' time on task is as reliable as, but faster than card sorting.

Raters' classification of learning tasks belonging to the intermediate categories were not always in agreement with the conceptual frame of reference. Compared to the extremes, the raters were also less confident about their ratings. If deviations occur between raters' classifications and our conceptual frame of reference, the deviations are mostly for tasks in the intermediate categories and seldom for the extremes. The deviation is never larger than one category.

There are four possible explanations for these observations. First, the size of a task (amount of cognitive operations) versus the sort of cognitive operations (level) makes it impossible to operationalize complexity in the same way for categories 2 and 3. As a result, and

contrary to our assumptions, categories 2 and 3 are not disjunctive but show some overlap. Second, the tasks included in the experiments were not unmistakably representative for our conceptual frame of reference because they were taken from existing instructional materials. Therefore, we cannot be sure that the anchor tasks were fully representative for their class. This probably had the largest effect on categories 2 and 3 because the attribution of intermediate categories of a scale is always more difficult than the attribution of categories at both ends (P.G. Swanborn, personal communication, December 4, 2001), especially if the complexity of learning tasks has a normal distribution. Third, the conceptual frame of reference was quite abstract and therefore the anchor tasks might have been too limited (the conceptual frame of reference and the function of the anchor tasks were not made explicit to the raters). Again, this may have had the largest effect on categories 2 and 3. Finally, since complexity is a multidimensional concept, our one-dimensional approach that concentrated only on intellectual operations might explain these small classification anomalies. Below, the four possible explanations will be treated in more detail.

The first reason for classification problems is that the way complexity is made operational for the categories 2 and 3 is contaminated by the size of a task (i.e., the number of cognitive operations that need to be carried out) versus the type of cognitive operations (i.e., the level of the cognitive operations to be carried out). As a result, complexity categories 2 and 3 do not perfectly match cognitive operation levels 2 and 3. The classes in our conceptual frame of reference for intellectual operations are expected to be disjunctive and therefore not to show overlap. Although we used Merrill's theory (1987), several other models for classifying cognitive operations show the same four categories (Anderson & Krathwohl, 2001; De Block, 1975; Crombag et al., 1979; Lewy & Báthory, 1994). Nevertheless, these categories do not specifically

represent levels of complexity. It is known from other models for determining task complexity that it is impossible to objectively weigh identified task characteristics (Campbell, 1988; Wood, 1986). This may also be the case for the categories in our conceptual frame of reference. The categories can be clearly identified, but they are not completely in line with increasing complexity. This is especially true for tasks belonging to category 2 ("understand a generality") and category 3 ("use"). In line with this is that the size of a task within these two categories of intellectual operations can vary enormously. Thus, in some circumstances understanding a generality (e.g., understanding the concept Justice) can be more difficult than using a particular piece of knowledge (e.g., applying a simple procedure for finding the maximum punishment that can be applied for a certain crime). In other words, a large task requiring low level cognitive operations might be more complex than a small task requiring higher level cognitive operations.

The second reason for classification problems, especially for category 2 and 3 tasks, is that the to-be-rated tasks were not specifically developed for the experiments and were therefore not unmistakably representative for our conceptual frame of reference. In addition, this may explain the marginal effect of the use of anchor tasks on the variance of the rating results, since the anchor tasks, ipso facto, also were not specifically developed as a representative task for a particular class. Therefore, it cannot be ruled out that they were an ambivalent representative of one class, especially since participants in Experiment 2 did not fully agree on the anchor tasks chosen for Experiment 3. In other words, the anchor tasks were not always perceived of as good representatives for their categories. Furthermore, there were no explanations given as to why particular anchor tasks represented a particular category. Together, these considerations may explain the limited value of the anchor tasks.

The third explanation for the classification problems is that we chose to use only one anchor task to represent a category. The use of only one anchor task might have resulted in an overly limited category representation. Indeed, participants indicated for each category that the anchor task representing this category was not of much help to them during their classifications. The idea that anchor tasks could have been too limited is also supported by the data of participants' rating criteria. Despite a close resemblance of the participants' criteria for rating the tasks with our conceptual frame of reference, the raters nevertheless show a high variance in the ratings for the separate tasks. Since the conceptual frame of reference was not made explicit to the participants, the function of the anchor tasks might also have been veiled to them and therefore too abstract. Anchor tasks showing the upper and lower limits of a category might have been more helpful.

The fourth reason for classification problems might be that we chose to operationalize complexity as a one-dimensional concept, namely on the dimension *intellectual operations*. As is the case for intelligence, complexity can also be regarded as a multidimensional concept. Guilford (1982) proposed a factor-analytic model of intelligence consisting of 150 independent abilities that result from the interaction of five types of contents, five types of operations, and six types of products. Sternberg (1985) went "Beyond IQ" offering a "triarchic theory of human intelligence" with three components: analytic (academic) intelligence, creative intelligence, and practical intelligence. Yet the fact that it is commonly accepted that intelligence is a multidimensional concept, this does not preclude research on any one of those dimensions. Complexity too is probably a multidimensional concept, with other dimensions being (i) quantity of information searched for and combined (more quantity is more complex) (ii) field or discipline (some disciplines are more complex than others), (iii) symbolic system of task-

formulation (text, and/or graphics, animations; some 'languages' and/or symbol systems are more complex than others), (iv) preferences and styles of the receiver (some people are text orientes, other iconic), et cetera.. Since intellectual operations appear to be hierarchical causing their relative contribution to a multidimensional complexity construct increases in the higher categories, we expect intellectual operations to be an important dimension in determining objective task complexity. In this, we lean on the work of important theorists in the field such as Bloom (see e.g., Anderson & Krathwohl, 2001) and Merrill (see, e.g., Merrill, 1987). However, more research is needed in which a multidimensional theoretical construct on learning task complexity should be tested.

Two promising lines for further research can be distinguished for measuring the complexity of learning tasks. First, research should be conducted as to whether the explanation of the conceptual frame of reference together with an explanation of the way that anchor tasks fit within this frame, would result in more confident ratings for the separate tasks. In the present study, the conceptual frame of reference was not made explicit, because it was yet unknown if participants' ratings would be similar to this frame. We now know that participants use similar criteria. Explaining the conceptual frame of reference in future work could result in more consensus between ratings for the separate tasks.

A second line of further research includes implementing the current instrument for measuring task complexity in an Instructional Design model to find out whether this results in improved, more effective designs and products. Nadolski, Kirschner, van Merriënboer & Hummel (2001; see also van Merriënboer, 1997) describe an ID-model whose application reduces the complexity of learning tasks in competency-based learning environments through a multiple-step whole-task approach, while not sacrificing the authenticity of the learning

experience. Figure 1 presents this two-phase, six-step ID-model. Phase 1 deals with Cognitive Task Analysis (steps 1 through 3) and Phase 2 with actual training design (steps 4 through 6), resulting in a detailed blueprint for the learning environment.

****

Insert Figure 1 about here

****

The six steps are: (1) decomposing the complex skill, (2) determining task complexity, (3) identifying Systematic Approaches to Problem solving (SAPs), (4) sequencing problems on the micro-level, (5) choosing problem formats, and (6) choosing the step size of SAPs that will be presented to learners as process worksheets (i.e., lists of prompting questions). Steps 2 and 6 involve the determination of task complexity. Up till now, teachers had to use a subjective, intuitive measure of task complexity because there was no available instrument for objectively measuring task complexity. But now that this instrument has been developed, experiments will be conducted to determine the optimal step size in process worksheets.

From a practical point of view, it is important to develop instruments for the measurement of task complexity in other domains. The conceptual frame of reference presented in this article offers a good starting point because the identified intellectual operations are not exclusive to the domain of Law. It should be stressed once again that an effective use of such an instrument presupposes that students confronted with the rated learning tasks have roughly the same prior knowledge. Indeed, an instrument for measuring task complexity has only limited value if students differ greatly in prior knowledge. A current trend in education is to develop personalized, student-centered instruction that takes differences in prior knowledge into account. Then, at first sight, it seems to be of little use to develop instruments that measure task

complexity in advance, that is, before the learning tasks are actually presented to the learners. Collecting data on subjective task complexity from learners then seems a workable solution for tailoring the instructional material "on the fly".

Research on instruments for measuring learning task complexity will become increasingly important because the demand for competency-based learning environments and CMPs is still growing. In such environments (e.g., Wöretshofer et al., 2000), it is of utmost importance to carefully adjust the complexity of learning tasks and the step-size of SAPs that support learners in performing those tasks to the target learners. Measurement instruments for task complexity support instructional designers in this process, yielding better support for learners and more effective learning.

References

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). A taxonomy for learning, teaching, and assessing. A revision of Bloom's taxonomy of eductional objectives. Addison Wesley Longman, Inc.

Bandura, A. (1982). Self-efficacy mechanism in human accuracy. American Psychologist, 37, 122-147.

Boggs, D. H., & Simon, R. J. (1968). Differential effect of noise on tasks of varying complexity. Journal of Applied Psychology, 52, 148-153.

Bonner, S. (1994). A model of the effects of audit task complexity. Accounting, Organizations and Society, 19(3), 213-234.

Brandle (2002). Sentence complexity [On-line]. Available: http://www.brandle.com.au/sentence.htm

Brown, J. S., Collins, A., & Duguid, S. (1989). Situated cognition and the culture of learning. Educational Researcher, 18, 32-42.

Burtch, L. D., Lipscomb, M. S., & Wissman, D. J. (1982). Aptitude requirements based on task difficulty (Report No. AFHRL-TR-81-34). Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower and Personnel Division.

Byström, K., & Järvelin, K. (1995). Task complexity affects information seeking and use. Information Processing & Management, 31, 191-213.

Campbell, D. J. (1988). Task complexity: a review and analysis. Academy of Management Review, 13, 40-52.

Campbell, D. J., & Gingrich, K. (1986). The interactive effects of task complexity and participation on task performance: A field experiment. Organizational Behaviour and Human Decision Processes, 38, 162-180.

Clough, P. (2000). Analysing style-readability [On-line]. Available: http://www.dcs.sheef.ac.uk/~cloughie/papers/readability.pdf

Crombag, H. F., Chang, T. M., Drift, K. D. J. M. van der, & Moonen, J. M. (1979). Onderwijsmiddelen voor de Open Universiteit Nederland: Functies en kosten [Eductional materials for the Open University of the Netherlands: Functions and costs]. Ministry of Education and Sciences, The Hague.

De Block, A. (1975). Taxonomie van leerdoelen [Taxonomy of learning objectives]. Standard Academic Publishers, Antwerp-Amsterdam.

Early, P. C. (1985). Influence of information, choice and task complexity upon goal acceptance, performance, and personal goals. Journal of Applied Psychology, 70, 481-491.

Flesch, R. (2003). Chapter 2: Let's Start With the Formula. In: How to write plain English [On-line]. Available: http://www.mang.canterbury.ac.nz/courseinfo/AcademicWriting/Flesch.htm

Guilford, J. P. (1982). Cognitive psychology’s ambiguities: Some suggested remedies. Psychological Review, 89, 48-59.

Hays, W. L. (1981). Statistics (3 rd ed.). New York: CBS College Publishing.

Huber, V. L. (1985). Effects of task difficulty, goal setting, and strategy on performance of a heuristic task. Journal of Applied Psychology, 70, 492-504.

Kernan, M. C., Bruning , M. S., & Miller-Guhde, L. (1994). Individual and group performance: Effects of task complexity and information. Human Performance, 7, 273-289.

Lewy, A., & Báthory, Z. (1994). The taxonomy of educational objectives in continental Europe, the Mediterranean and the Middle East. In L.W. Anderson & L.A. Sosniak (Eds.), Bloom's taxonomy: A fourty-year retrospective (pp. 146-163). The University of Chicago Press. Chicago, Illinois.

Maynard, D. C., & Hakel, M. D. (1997). Effects of objective and subjective task complexity on performance. Human Performance, 10(4), 303-330.

Merrill, M. D. (1983). Component display theory. In C. M. Reigeluth (Ed.), Instructional design theories and models: An overview of their current status (pp. 278-333). Hillsdale, NJ: Lawrence Erlbaum.

Merrill, M. D. (1987). The new component design theory: Instructional design for courseware authoring. Instructional Science, 16, 19-34.

Nadolski, R. J., Kirschner, P. A., van Merriënboer, J. J. G., & Hummel, H. G. K. (2001). A model for optimizing step size of learning tasks in competency-based multimedia practicals. Educational Technology Research and Development, 49, 87-103.

Paas, F. G. W. C., & van Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem solving skills: A cognitive load approach. Journal of Educational Psychology, 86, 122-133.

Pikulski, J. J. (2002). Readability [On-line]. Available: htttp://www.eduplace.com/state/author/pikulski.pdf

Scott, W. E., Fahr, J., & Podsakoff, P. M. (1988). The effects of "intrinsic" and "extrinsic" reinforcement contingencies on task behavior. Organizational Behavior and Human Decision Processes, 41, 405-425.

Siegel, S. (1956). Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill, Inc.

Sternberg, R. J. (1985). Beyond IQ: A triarchic theory of human intelligence. New York: Cambridge University Press.

Taylor, M. S., (1981). The motivational effects of task challenge: A laboratory investigation. Organizational Behavior and Human Performance, 27, 255-278.

Van Merriënboer, J. J. G. (1997). Training complex cognitive skills. Englewood Cliffs, New Jersey: Educational Technology Publications.

Vaso, G. (2000). Determining the Readability of a Book. Eclectic Homeschool. Bellevue [On-line]. Available: http://www.eho.org/pdf/100ln.pdf

Westera, W., & Sloep, P. B. (1998). The Virtual Company: Toward a self-directed, competence-based learning environment in distance education. Educational Technology, 38, 32-37.

Winne, P. H. (1997). Experimenting to bootstrap Self-Regulated Learning. Journal of Educational Psychology, 89, 397-410.

Wood, R. E. (1986). Task complexity: Definition of the construct. Organizational behaviour and human decision processes, 37, 60-82.

Wöretshofer, J., Nadolski, R. J., Starren-Weijenberg, A. M. A. G., Quanjel-Schreurs, R. A. M., Aretz, C. C.W. M., Meer, N. H. W. van der , Martyn, G., Brink, H. J. van den, Slootmaker, A., & Berkhout, J. (2000). Preparing a Plea [Pleit voorbereid] (version 1.0) [multimedia CD-ROM]. Heerlen, The Netherlands: CIHO.

Table 1

Concordance coefficient for criminal and civil law tasks by participants' area of expertise

	Criminal law tasks		Civil law tasks	
Participants' expertise	4-points	16-points	4-points	16-points
Criminal law (n = 7)	.547	.570	.518	.592
Civil law (n = 12)	.601	.651	.540	.613

Note. p < .01 for all values for Kendalls W (concordance coefficient).

Table 2

Ratings for criminal and civil law learning tasks (4 points-ranking scale) and student's time-on-task statistics

Criminal law tasks									Civil law tasks								
card sort						time on task (min)			card sort						time on task (min)		
Id	M	SD	Mdn	Rm,Rf	P(c=ci)	M	SD	rm	Id	M	SD	Mdn	Rm,Rf	P(c=ci)	M	SD	rm
cr14	1.32	.58	c1	c1,c1	.72	13	8	c1	ci8	1.21	.42	c1	c1,c1	.86	11	5	c1
cr8	1.37	.83	c1	c1,c1	.64	10	6	c1	ci13	1.26	.65	c1	c1,c1	.73	9	10	c1
cr9	1.37	.68	c1	c1,c1	.67	12	4	c1	ci7	1.37	.49	c1	c1!,c2	.73	13	6	c1!
cr15	1.89	.94	c2	c2!,c1	.35	13	10	c1	ci10	1.58	.69	c1	c1,c1	.55	20	25	c2!
cr6	1.89	.88	c2	c2!,c3	.36	19	14	c2!	ci11	2.06	1.03	c2	c2!,c1	.32	16	12	c1
cr2	2.16	.96	c2	c2!,c3	.34	31	52	c3	ci15	2.21	.86	c2	c2,c2	.37	23	24	c2
cr5	2.21	.79	c2	c2,c2	.40	20	13	c2	ci5	2.21	.86	c2	c2!,c3	.37	21	15	c2!
cr11	2.26	.73	c2	c2,c2	.42	19	10	c2	ci1	2.53	.70	c3	c3!,c2	.39	19	12	c2
cr4	2.32	1.00	c2	c2!,c3	.32	27	26	c3	ci16	2.58	1.02	c3	c3!,c2	.30	23	26	c3!
cr12	2.68	.75	c3	c3!,c2	.40	23	12	c3!	ci14	2.95	.97	c3	c3!,c4	.33	32	38	c3!

Table 2 (continued)

cr10	2.84	.60	c3	c3!,c2	.51	25	11	c3!	ci9	3.00	1.05	c3	c3!,c4	.31	134	180	c4
cr16	2.89	.88	c3	c3,c3	.36	21	12	c2!	ci2	3.32	.67	c3	c3,c3	.39	37	36	c4!
cr13	3.53	.84	c4	c4,c4	.59	141	368	c4	ci4	3.32	.75	c3	c3!,c4	.37	36	33	c3!
cr7	3.63	.68	c4	c4,c4	.67	56	63	c4	ci12	3.42	.77	c4	c4!,c3	.55	36	32	c3
cr1	3.68	.58	c4	c4,c4	.73	128	281	c4	ci3	3.42	.77	c4	c4,c4	.55	45	42	c4
cr3	3.84	.37	c4	c4,c4	.92	44	36	c4	ci6	3.53	.77	c4	c4!,c3	.60	43	45	c4!

Notes. Id = identification for the task, Rm = rank based on the mean ranking-score of the card sort, Rf = rank based on conceptual frame of reference, rm = rank based on the mean ranking-score of the estimated students' time on task, != deviation from conceptual frame of reference. For card sort: c1 = very simple task [1, 1.66], c2 = simple task (1.66, 2.5], c3 = complex task (2.5, 3.33], c4 = very complex task (3.33, 4]. P(c=ci); confidence 'ci' is correct, ci is based on the mean ranking score of the card sort.

Table 3

Ratings for criminal and civil law learning tasks (16 points-ranking scale)

Criminal law tasks			Civil law tasks		
Id	RCS	RTm	Id	RCS	RTm
cr14	1	4	ci8	1	2
cr9	3	2	ci13	2	1
cr8	2	1	ci7	3	3
cr15	5	3	ci10	4	6
cr11	7	6	ci11	5	4
cr6	4	5	ci16	8	9
cr2	6	12	ci15	7	8
cr4	9	11	ci1	9	5
cr5	8	7	ci5	6	7
cr12	10	9	ci14	11	10
cr16	11	8	ci4	12	11
cr13	13	16	ci9	10	16
cr10	12	10	ci2	14	13
cr7	14	14	ci12	13	12
cr3	16	13	ci6	16	14
cr1	15	15	ci3	15	15

Notes. RCS = rank card sort derived from mean rank in Kendalls W test. RTm = rank time on task derived from mean time on task

Table 4

Ratings for law learning tasks in Experiments 2 and 3 (4 points-ranking scale)

Law tasks (Experiment 2)						(same tasks in Experiment 3)											
						Teachers						Students					
Id	M	SD	Mdn	Rm,Rf	P(c=ci)	M	SD	Mdn	Rm, Rf	P(c=ci)	rank	M	SD	Mdn	Rm,Rf	P(c=ci)	rank
t1	1	0	c1	c1,c1	1	1	0	c1	c1, c1	1	1	1	0	c1	c1, c1	1	1
t10	1.08	.29	c1	c1,c1	.98	anchor task											
t9	1.17	.39	c1	c1,c1	.90	1.31	.75	c1	c1, c1	.68	3	1.17	.39	c1	c1, c1	.91	2
t12	1.42	.51	c1	c1,c1	.68	1.15	.55	c1	c1, c1	.82	2	1.17	.39	c1	c1, c1	.91	3
t20	1.42	.51	c1	c1,c1	.68	1.46	.66	c1	c1, c1	.62	4	1.33	.65	c1	c1, c1	.70	4
t8	1.67	.65	c2	c2,c2	.41	1.62	.65	c1	c1!, c2	.52	5	1.33	.65	c1	c1!, c2	.70	5
t18	1.92	.67	c2	c2!,c3	.30	1.92	.64	c2	c2!, c3	.48	6	2.17	.72	c2	c2!, c3	.44	6
t23	2.08	1.08	c2	c2!,c1	.30	2.54	1.05	c2*	c3!, c1	.29	10	2.25	.97	c2	c2!, c1	.33	7
t22	2.17	1.03	c2	c2,c2	.32	2.54	1.13	c3	c3!, c2	.27	11	2.42	.79	c2	c2, c2	.37	10
t21	2.25	.75	c2	c2!,c3	.41	2.62	.65	c3	c3, c3	.44	13	2.75	.87	c3	c3, c3	.36	12
t5	2.33	.78	c2	c2, c2	.39	anchor task											
t11	2.42	.79	c2	c2,c2	.37	2.08	.64	c2	c2, c2	.49	7	2.25	.97	c2	c2, c2	.33	8

Table 4 (continued)

t14	2.50	.90	c2	c2!,c3	.32	2.15	.80	c2	c2!, c3	.40	8	2.33	.65	c2	c2, c2	.45	9
t15	2.67	.49	c3	c3!,c2	.55	2.46	.78	c2	c2, c2	.37	9	2.50	.67	c2	c2, c2	.39	11
t6	2.67	.78	c3	c3,c3	.39	2.77	.44	c3	c3, c3	.50	14	2.83	.72	c3	c3, c3	.43	13
t17	3.00	.60	c3	c3,c3	.51	anchor task											
t19	3.00	1.13	c3	c3!,c4	.28	3.54	.88	c4	c4, c4	.59	16	3.42	.67	c4	c4, c4	.55	16
t7	3.17	.72	c3	c3!,c2	.41	2.62	.87	c3	c3!, c2	.35	12	2.92	.79	c3	c3!, c2	.40	14
t24	3.41	.51	c3*	c4,c4	.56	3.92	.28	c4	c4, c4	.98	20	3.50	.80	c4	c4, c4	.58	17
t13	3.5	.90	c4	c4,c4	.58	3.69	.63	c4	c4, c4	.72	18	3.75	.62	c4	c4, c4	.75	18
t3	3.58	.51	c4	c4!,c3	.69	3	.82	c3	c3, c3	.38	15	3.17	.58	c3	c3, c3	.49	15
t4	3.75	.45	c4	c4,c4	.82	3.54	.52	c4	c4, c4	.66	17	3.92	.29	c4	c4, c4	.98	20
t16	3.92	.29	c4	c4,c4	.98	3.77	.44	c4	c4, c4	.84	19	3.83	.39	c4	c4, c4	.90	19
t2	3.92	.29	c4	c4,c4	.98	anchor task											

Notes. Id = identification for the task, Rm = rank based on the mean ranking-score of the card sort, Rf = rank based on conceptual frame of reference, != difference with conceptual frame of reference, * = difference between classifications based on mean or median. c1 = very simple task [1, 1.66], c2 = simple task (1.66, 2.5], c3 = complex task (2.5, 3.33], c4 = very complex task (3.33, 4]. P(c=ci); confidence 'ci' is correct, ci is based on the mean ranking score of the card sort

Table 5

Usefulness of anchor tasks for rating

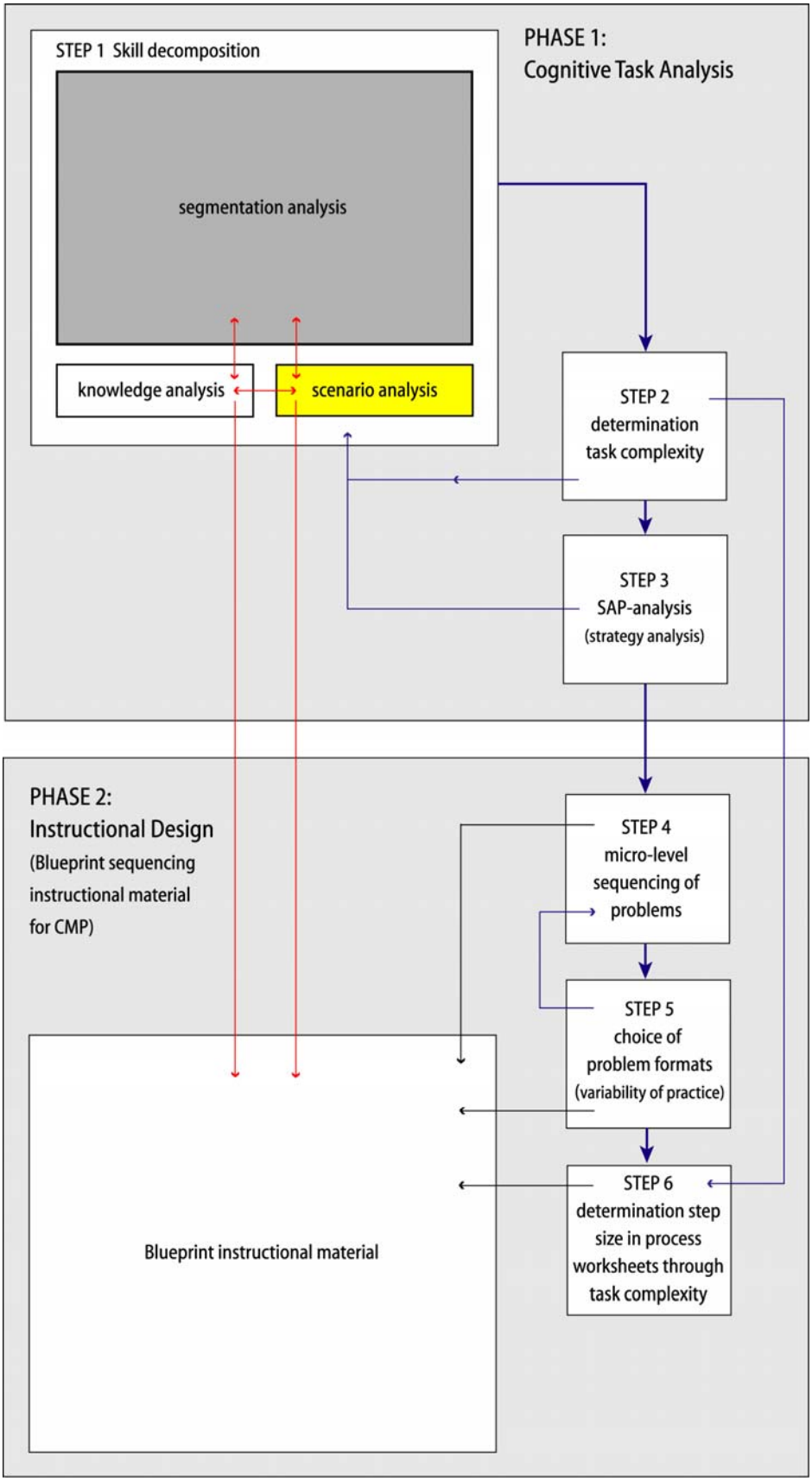
	teachers			students		
	M	SD	Mdn	M	SD	Mdn
a1	2.80	1.40	3	4.08	1.31	5
a2	2.55	1.37	2	3.25	1.06	3
a3	2.64	1.50	2	3.25	.87	3
a4	2.46	1.29	2	3.58	1.38	4

Notes. a1 = anchor task for category 1 (very simple), a2 = anchor task for category 2 (simple), a3 = anchor task for category 3 (complex), a4 = anchor task for category 4 (very complex). Participants indicated on a 6-point categorical scale their agreement (1 = totally disagree, 6 = totally agree) with the assertion "anchor task a[i] was useful for the classification of the tasks in category [i]"([i] = 1, 2, 3 or 4).

Figure captions

Figure 1. The two-phase, six-step ID-model for CMP-development.

Figure 1. The two-phase, six-step ID-model for CMP-development.



Appendix

Examples of to be rated learning tasks

very simple	simple	complex	very complex
Task formulation (to student): After celebrating in Amsterdam John is rather drunk and aggressive. As John is on his way home on December 12th 1995, someone gets in his way. He hits this person, Tim, hard on his nose. A while later, he runs into his old neighbour, who is also named Tim and this Tim also gets punched in the nose by John. Only the first Tim reports the assault to the police. John is charged as follows: "On or about December 12th in Amsterdam, a person named Tim was assaulted by John by deliberately punching him in the nose. As a result he was painfully hurt". Presented as evidence were a police report with the testimony of the (first) Tim and a testimony of a doctor, saying that after his examination on December 12th 1995, Tim was found to have a broken nose. John appears in court and states that he has no idea which fact the Prosecutor is referring to. Question: What decision should be made by the judge? a. declare the subpoena invalid, because it does not meet the criteria of art. 261 Sv. b. acquit John as the charge is unclear and cannot be proven. c. discharge John, as the fact cannot be qualified. d. declare him guilty of charge if no exclusion of punishment presents itself.	Task formulation (to student): The Dutch trawler fleet consists of sixteen trawlers. They are fishing for herring, mackerel, and horsmackerel in the North Sea, but especially in more remote fishery grounds. This is regarded as 'big' seafishing, a healthy business which seldom attracts publicity. However, a short time ago a fight occurred on a Dutch traylor in the territorial waters of Denmark. In this fight, a member of the crew was badly injured by an oriental handweapon. The prosecutor prosecutes the suspect on the charge of illegal possession of arms and attempted manslaughter, at least assault and battery. The suspect declares at the trial that he cannot be accused for this crime in the Netherlands, but he should stand trial in Denmark instead. Question: Is the suspect correct or not? Motivate your answer.	Task formulation (to student): During a football match between Ajax and Feyenoord, a Feyenoord player, Sjaak, is so irritated by the behavior of some Ajax-supporters, that he takes off his shoe and throws it at the supporters. A cleat on his shoe unfortunately hits Bram, one of the supporters, in the eye, damaging the retina. Sjaak is ejected and Bram needs treatment by his family doctor. He sends Bram to the hospital, to see an ophtalmologist. An operation is needed but there is a waiting list in the hospital which delays the operation for two months. Bram has been given instructions to visit the hospital on an empty stomach. In spite of this, Bram has eaten breakfast on the day of the operation causing a problem with the anesthesia. Bram dies during the operation. Question: Motivate the chances of the prosecutor in prosecuting Sjaak.	Task formulation (to student): May 1999. The 42-year old Georgette Smith is paralysed from the neck down ever since her 68-year old mother shot her in March. Her mother, blind in one eye, was angry because she heard that her daughter was planning to move her to a home for the elderly. Because of this anger, she shot her daughter and attempted to shoot her daughter's boyfriend. The bullet hit the spiral cord causing the paralysis. She cannot swallow anymore and has to be fed artificially. According to her doctors, her situation is irriversible. She has received a judges permission to be allowed to die. Therefore, it is possible that her mother can be prosecuted for manslaughter. Suppose the situation above has occurred in the Netherlands, and the victim uses her right to die by letting the medical apparatus be turned off. Question: Describe the chances of prosecuting the mother for manslaughter and also attend to arguments which could provide a conviction.

Art. 261 Sv refers to an article in Dutch Criminal Law. Answers are left out in these examples because of their substantive length. Original formulations were all in Dutch.