Psychological Test and Assessment Modeling, Volume 56, 2014 (1), 83-103

The systematic variation of task characteristics facilitates the understanding of task difficulty: A cognitive diagnostic modeling approach to complex problem solving

Samuel Greiff[1], *Katarina Krkovic*[2] *& Gabriel Nagy*[3]

Abstract

Since the 1960ies, when pioneering research on Item Response Theory (IRT) was published, considerable progress has been made with regard to the psychometrical quality of psychological assessment tools. One recent development building upon IRT is the introduction of Cognitive Diagnostic Modeling (CDM). The major goal of introducing CDM was to develop methods that allow for examining which cognitive processes are involved when a person is working on a specific assessment task. More precisely, CDM enables researchers to investigate whether assumed task characteristics drive item difficulty and, thus, person ability parameters. This may – at least according to the assumption inherent in CDM - allow conclusions about cognitive processes involved in assessment tasks. In this study, out of the numerous CDMs available the Least Square Distance Method (LSDM; Dimitrov, 2012) was applied to investigate psychometrical qualities of an assessment instrument measuring Complex Problem Solving (CPS) skills. For the purpose of the study, two task characteristics essential for mastering CPS tasks were identified ex-ante – degree of connectivity and presence of indirect effects by adding eigendynamics to the task. The study examined whether and how the two hypothesized task characteristics drive item difficulty of two CPS dimensions, knowledge acquisition and knowledge application. The sample consisted of 490 German high school students, who completed the computer-based CPS assessment instrument MicroDYN. The two task characteristics in MicroDYN items were varied systematically. Results obtained in

[1] University of Luxembourg

[2] *Correspondence concerning this article should be addressed to:* Katarina Krkovic, PhD, EMACS unit, University of Luxembourg, 6, rue Richard Coudenhove Kalergi, 1359 Luxembourg-Kirchberg, Luxembourg; email: katarina.krkovic@uni.lu

[3] Leibniz Institute for Science and Mathematics Education

LSDM indicated that the two hypothesized task characteristics, degree of connectivity and introducing indirect effects, drove item difficulty only for knowledge acquisition. Hence, other task characteristics that may determine item difficulty of knowledge application need to be investigated in future studies in order to provide a sound measurement of CPS.

Key words: item response theory; cognitive diagnostic modeling; least square distance method; complex problem solving; task characteristics

Educational systems and the labour market are not static phenomena – they change and evolve over time. In line with this fact, the last decades have been marked by a notable shift from routine tasks (i.e., tasks requiring recurring, monotonous actions) to non-routine tasks (i.e., tasks requiring complex, dynamic actions) in various areas of everyday life. As a result, complex, so-called transversal skills are necessary to successfully master non-routine tasks and, thus, exponentially gain in importance (Autor, Levy, & Murnane, 2003). Moreover, the acquisition of such complex skills is becoming an essential part of many educational curricula (Mayer & Wittrock, 2006). Consequently, assessments of transversal skills – Creativity, Complex Problem Solving, Computer Literacy, and Collaboration (i.e. 21st century skills; Binkley et al., 2012) – are included in a number of educational large-scale assessments. For instance, a computer-based assessment of Complex Problem Solving (CPS) was employed in the arguably most comprehensive international large-scale assessment, the Programme for International Student Assessment (PISA), in its 2012 survey (OECD, 2013).

However, in order to provide a sound measurement of any construct an extensive analysis of the assessment tool and its psychometric characteristics is warranted. To this end, recent developments within the scope of Item Response Theory (IRT) offer new ways of examining psychometric qualities of assessment tools. In this study, we use an advanced statistical method following IRT to profound our understanding of CPS and to assure the validity of an instrument used to assess CPS. More specifically, with the help of Cognitive Diagnostic Models[4](CDM; Rupp & Templin, 2008) we examine if and how different task characteristics drive CPS item difficulties. This may further help us understand, which underlying cognitive processes are involved in solving CPS assessment tasks and which processes are mastered at different ability levels.

Complex Problems Solving

Complex problems are characterized by their interactivity and dynamics (Wirth & Klieme, 2003). They require the problem solver to actively investigate the problem in order to acquire the information necessary to solve it (the interactive aspect). Further, the

[4] Please note that different authors use also following terms to refer to CDM: cognitive psychometric models, multiple classification models, latent response models, restricted latent class models, structured located latent class models, or structured IRT models (as reviewed by Rupp and Templin, 2008).

problem changes as a result of user interaction and/or time (the dynamic aspect) (cf. Buchner, 1995; Funke, 2001). An often-described CPS task is handling an MP3 player for the first time ever. A person who has no previous knowledge of an MP3 player needs to first explore the object, try out different commands. In doing so, a person acquires knowledge on how the system functions. After acquiring knowledge, a person has to apply this knowledge in order to reach specific goals, for example, making a playlist that contains only a limited number of songs (Funke, 2001; Greiff, 2012a). Hence, in theory CPS is composed of two dimensions – knowledge acquisition and knowledge application (cf. Mayer & Wittrock, 2006; Novick & Bassok, 2005). Furthermore, these two dimensions of CPS have been proven as separable in several empirical studies (e.g., Wüstenberg, Greiff, & Funke, 2012; Greiff et al., 2013).

Assessment of complex problem solving

Researchers early recognized the importance of CPS skills and their assessment has been an appealing research topic in the past decades. However, up to now there is little reliable information available about cognitive processes that take place when solving a complex problem. However, new assessment approaches combined with innovative statistical methods, such as CDM may allow us to investigate this research question. More specifically, earlier CPS assessment approaches had many constraints, and only recently new assessment instruments allow a reliable and valid assessment. First measurement attempts were based on laboratory computer simulations trying to imitate real life problem situations such as managing of a tailorshop, or of an entire city (Dörner, 1980, 1986). Such simulations enabled researchers to investigate the limits of human capability in managing complexity (cf. Dörner, 1980). However, aforementioned experimental operationalizations had severe limitations, not allowing their application for the assessment of individual skills outside of laboratory setting. Some of these limitations were the unacceptably long testing time, content of the tasks that substantially relied on previous knowledge, inadequate scoring of tasks, or an unsatisfying internal validity of instruments (cf. Greiff, 2012b). The introduction of formal frameworks (Funke, 2001) was a first reaction to measurement issues associated with early CPS tasks. Their implementation as an assessment approach for measuring CPS was an important step forward in developing a sound measurement tool. Formal frameworks aim at systematically constructing CPS tasks and at describing their underlying structure independent of semantic embedment. That is, problem solvers are confronted with a task composed of precisely defined components. Subsequently, this task is layered with an arbitrarily chosen semantic cover.

One particular framework, Linear Structural Equation (LSE) systems, has been widely perceived by CPS research and led to the development of a considerable amount of tasks, such as Multiflux (Kröner, Plass, & Leutner, 2005), Genetics Lab (Sonnleitner et al., 2012), or ColorSim (Kluge, 2008). In LSE systems such as the one depicted in Figure 1, input variables (in Figure 1: X_1, X_2, X_3) have an impact on output variables (in Figure 1: Y_1, Y_2, Y_3). Generally, a person can only manipulate the input variables (cf. Funke, 2001; Greiff & Funke, 2010). Further, output variables may be related to each other, which is labeled as a side effect (in Figure 1: Y_2 to Y_3). Another possibility is that an output variable is related to itself and changes independently from other influences, which is labeled as an eigendynamic (in Figure 1: Y_1). Direct connections from input to output variables are labeled as direct effects (i.e., relations from X_n to Y_n), whereas connections between output variables (i.e., side effects, eigendynamics) are labeled as indirect effects (cf. Greiff, Wüstenberg, & Funke, 2012).

The number of equations necessary to describe an entire LSE system is equal to the number of output variables, which are denoted by Y. For the LSE system example provided by Greiff et al. (2012; cf. Figure 1), the following equations are required:

$$Y_1(t+1) = a_1 * X_1(t) + a_2 * Y_1(t) \tag{1}$$

$$Y_2(t+1) = a_3 * X_2(t) + Y_2(t) \tag{2}$$

$$Y_3(t+1) = a_4 * X_2(t) + a_5 * X_3\,(t) + a_6 * Y_2(t) + Y_3(t) \tag{3}$$

with t = discrete time steps, a_i = arbitrary path coefficients, $a_i \neq 0$, and .. ≠ 1.

The advantage of LSE systems is that due to their dependency on linear equations they allow for variation of task characteristics, which gives researchers the means to systematically adjust a wide range of task difficulties in an assessment of CPS. Hence, with the help of CDM, we can analyze which ability level is necessary to master a specific task characteristic. This, in turn, may enable us to investigate the nature of cognitive processes involved in the CPS assessment and in the construct itself.

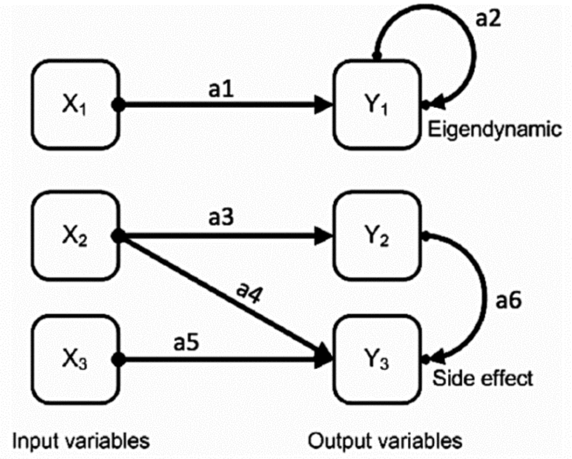

Figure 1:
Example of the structure of a typical LSE system displaying three input (X_1, X_2, X_3) and three output (Y_1, Y_2, Y_3) variables (Greiff et al., 2012).

Between- vs. within-component perspective

When investigating CPS or other cognitive constructs, there are two different perspectives for research. The common one is the between-component perspective. Here, the main goal is to investigate which components are relevant for CPS. However, the understanding of CPS should not stop at knowing which components are relevant (between-component perspective), but also which task characteristics of these components drive item difficulty and person ability estimates (within-component perspective). Therefore, it is not only important to differentiate between knowledge acquisition and knowledge application as two separate CPS dimensions, but also to identify which task characteristics determine the task difficulty within these two dimensions.

When constructing and validating an assessment instrument for CPS, it is therefore essential to address the question of the nature of processes problem solvers engage into when performing on specific components of a CPS task (Rupp & Templin, 2008). That is, in order to grasp the degree to which a problem solver can master a certain CPS component, a profound understanding of the cognitive processes contributing to this performance is mandatory (Wilhelm & Robitzsch, 2009). That is, research on CPS task characteristics can bring us one step nearer to discovering the nature of cognitive processes involved in the tasks.

For instance, within knowledge application – one of the two overarching problem solving processes besides knowledge acquisition – different tasks place different cognitive demands on a problem solver. Whereas some CPS tasks may be static and do not change without respondents' intervention, some may change dynamically over time. This rate of change over the course of time could be slow for some tasks (e.g., population growth), or rapid and abrupt for others (e.g., spread of diseases). The extent, to which respondents master different dynamics across tasks, indicates their ability level and gives an insight into the nature of cognitive processes involved in knowledge application.

In order to illustrate how the aforementioned within-component perspective enhances the understanding of underlying cognitive processes, we consider a classical example from mathematical problem solving. In his linear logistic test model (LLTM), Fischer (1973) explains difficulty of mathematic items through a set of eight basic task characteristics such as differentiating a polynomial or using the quotient rule. By reducing a large number of item difficulty estimates to considerably fewer attribute parameters a substantive understanding of which processes respondents apply to master items is obtained. In fact, many studies attempt to predict item difficulty by identifying task characteristics thereby gathering insights into the underlying cognitive processes (e.g., Buck & Tatsuoka, 1998; Green, 1984; Poinstingl, 2009). For instance, in order to determine item generating rules for the verbal reasoning test "Family Relation Reasoning Test" (FRRT), Poinstingl (2009) examines if specific task characteristics that were theoretically identified beforehand are able to explain item difficulty parameters obtained within a Rasch model. Using Fischer`s (1973) LLTM matrix of weights, Poinstingl shows that assumed relevant task characteristics are not entirely reflected in the empirically obtained data, and that, thus, the construct validity of the FRRT must be further investigated. As a result, Poinstingl (2009) concludes that other task characteristics must be taken into account in order to generate further items for a revised version of the FRRT. Generally speaking, discovering which task characteristics are determining the item difficulty allows researchers generating adequate items for all ability levels, thereby improving the psychometrical qualities of an assessment instrument.

Methods used in the enterprise of attributing item difficulty estimates to a smaller number of parameters, such as the LLTM matrix of weights used by Fischer (1973) and Poinstingl (2009), embrace a variety of different models ranging from classical multiple regressions to modern CDM. Particularly the latter methods can guide theory-based task construction and enhance a substantive understanding of CPS and the cognitive processes problem solvers engage into while solving a problem task. Hartig (2007) points out that hypotheses on task characteristics and accompanying cognitive processes, which essentially inform test development, are ideally addressed ex-ante by substantive and experimental research in the respective field. However, the systematic procedure of ex-ante manipulation of task characteristics, as the one in Poinstingl`s study, is hardly ever found. Even Fischer (1973) derived characteristics of his mathematical tasks in a post-hoc analysis. Importantly, LSE embedded assessments, with their systematic and linear structure make the ex-ante manipulation of task characteristics realistic and manageable. In the present study, we focus on the construct of CPS and aim at estimating item difficulty parameters of CPS tasks using a relatively new and valid assessment instrument, the MicroDYN approach, which is embedded in the LSE framework. In doing so, we apply ex-ante systematical variation of task characteristics, which are in theory assumed to be essential for the difficulty of CPS items.

Task characteristics and task difficulty in CPS

The LSE framework (cf. Figure 1), allows to vary discrete task characteristics and to observe their impact on performance. This may serve as a method to detect which task attributes impact item difficulty, and may further indicate, arguably, which cognitive processes are at work when solving an item.

So far, experimental research on CPS has provided empirical information on three task characteristics in LSE, thus allowing for ex-ante assumptions about item difficulty. First, the number of constituting elements (i.e., the number of variables the problem solver needs to take into account) can be varied. For example, the LSE system in Figure 1 consists of three input and three output variables that need to be considered. Second, degree of connectivity includes all relations between variables that a problem solver needs to consider. In Figure 1, connections can be recognized by arrows connecting output and input variables (i.e., direct effects; Funke 2001), arrows connecting two output variables (i.e., side effect; Funke 2001), or curved arrows that point from and back to the same output variable (i.e., eigendynamic; Funke 2001). Finally, the third task characteristic is the presence of indirect effects (i.e., side-effects, eigendynamics; Funke 2001) within an LSE system. The defining characteristic of an indirect effect is that it cannot be changed directly by influencing one input variable. The eigendynamic in the example in Figure 1 is presented by the arrow pointing from and back to the output variable Y_1. Also in Figure 1, a side-effect is presented by the arrow connecting Y_2 and Y_3. All in all, the example in Figure 1 has a degree of connectivity equal to 6 (four direct connections, one eigendynamic, and one side-effect). Because there is an eigendynamic present, it can be scored as 1 with regard to the presence of indirect effects (1 for the presence of eigendynamics or 0 for no eigendynamics in the task). Side-effects can be scored in the same way as the presence of eigendynamics, with 1 for the presence of a side-effect (or effects) and 0 for no side-effects.

In her study on CPS assessment, Kluge (2008) uses the task ColorSim embedded into the LSE framework, which allows task characteristics variation (cf. Table 1). She introduces three task difficulty levels for ColorSim tasks, depending on the two task characteristics – degree of connectivity and presence of indirect effects. Kluge (2008) reports that simultaneously increasing the degree of connectivity as well as introducing indirect effects in LSE considerably reduces the knowledge acquired and the achieved level of knowledge application. Hence, varying different task characteristics reflects directly on the task difficulty (cf. Table 1). However, in her study, Kluge (2008) varies different task characteristics simultaneously and unsystematically, making it impossible to draw reliable conclusions regarding the question, which task characteristics determine item difficulties. Specifically, from the linear equations provided in Table 1 follows that both task characteristics – degree of connectivity and presence of indirect effects – were varied simultaneously in order to achieve medium difficulty. The same was done to construct a difficult version of the ColorSIM task. Thus, it can be concluded that if combined, the two hypothesized task characteristics influence the difficulty of items. However, no statement can be made about how each task characteristic separately influences the item difficulty.

Table 1:
Linear equations of the three task versions with different difficulties of ColorSim, with eigendynamics (in bold) and side effects (in italics) (Kluge, 2008).

Task difficulty: Easy
$\text{Green}_{t+1} = 10{*}x_t$
$\text{Black}_{t+1} = 3{*}z_t + 1.0{*}\text{black}_t$
$\text{Yellow}_{t+1} = 2{*}y_t + 0.5{*}z_t$
Task difficulty: Medium
$\text{Green}_{t+1} = 10{*}x_t + 1.1{*}\mathbf{green_t}$
$\text{Black}_{t+1} = 3{*}z_t + 1.0{*}\text{black}_t + 0.2{*}\mathit{yellow}_t$
$\text{Yellow}_{t+1} = 2{*}y_t + 0.5{*}z_t$
Task difficulty: Difficult
$\text{Green}_{t+1} = 10{*}x_t + 1.1{*}\mathbf{green_t} + 0.5{*}\mathit{yellow}_t$
$\text{Black}_{t+1} = 3{*}z_t + 1.0{*}\text{black}_t + 0.2{*}\mathit{yellow}_t$
$\text{Yellow}_{t+1} = 2{*}y_t + 0.5{*}z_t + 0.9{*}\mathbf{yellow_t}$

Funke (1992) shows that performance on both dimensions – knowledge acquisition and knowledge application – drastically decreases when indirect effects are introduced. The difficulty problem solvers encounter when dealing with indirect effects is experimentally bolstered by Greiff (2012a), who shows how tasks become gradually more difficult when the number of indirect effects is increased. Additionally, the degree of connectivity is a major constituent of difficulty, whereas the number of constituting elements taken by itself is not (Greiff, 2012a).

In summary, experimental studies have produced substantial effects on task difficulty for the degree of connectivity and for indirect effects alluding to the increased cognitive demands associated with these task characteristics. More specifically, it is assumed that in line with Sweller (2003) by rising connectivity and by adding indirect effects intrinsic and germane cognitive load substantially increase leading to a larger percentage of errors. In fact, Sweller (2005) considers complexity of underlying task structure a major cause of cognitive load. In contrast, the number of constituting elements seems to be of little importance in LSE-based CPS tasks.

Research questions

The importance of determining task characteristics in order to understand a construct is widely accepted and has been a research question in several experimental studies (e.g., Fischer, 1973; Kluge, 2008; Poinstingl, 2009). However, up until now, there is no record of a study examining effects of task characteristics on CPS item difficulties by using CDM. Therefore, the main research goal of this study is to enhance our knowledge about how task characteristics drive item difficulty parameters of knowledge acquisition and knowledge application in CPS. To yield sufficient variation in difficulty, only the two task characteristics that have been experimentally shown to impact task difficulty before – degree of connectivity and presence of indirect effects (eigendynamics) – are of interest in our study.

Furthermore, in their empirical studies, Greiff et al. (2012) and Wüstenberg et al. (2012) report generally lower reliability and predictability of knowledge application. Therefore, it is assumed that effects of cognitive characteristics within LSE tasks on performance measures are similar in their nature for both CPS dimensions with effect sizes somewhat smaller for knowledge application. Overall, we derive three hypotheses.

Hypothesis 1:

Hypothesis 1.1: Increasing the degree of connectivity in a CPS task significantly augments difficulty of knowledge acquisition items.

Hypothesis 1.2: Introducing indirect effects (eigendynamics) to a CPS task significantly augments difficulty of knowledge acquisition items.

Hypothesis 2:

Hypothesis 2.1: Increasing the degree of connectivity in a CPS task significantly augments difficulty of knowledge application items.

Hypothesis 2.2: Introducing indirect effects (eigendynamics) to a CPS task significantly augments difficulty of knowledge application items.

Hypothesis 3:

Larger effect sizes are expected for varying task characteristics for knowledge acquisition items than for knowledge application items.

Material and methods

This study aims at elaborating the construct of CPS and its defining dimensions by discovering how different task characteristics drive item difficulty parameters in a MicroDYN assessment of CPS. Two task characteristics – degree of connectivity and presence of indirect effects – were systematically varied ex-ante. Subsequently, the MicroDYN assessment tool was administered to a sample of German students within secondary education. Finally, CDM was applied on the empirically obtained data in order to investigate whether the hypothesized task characteristics have an impact on item difficulty of two CPS dimensions – knowledge acquisition and knowledge application.

Participants

The sample was composed of N=490 students[5] (248 female; M_{Age}=15.8; SD_{Age}=2.0). Participants were tested at computer facilities of a school located in the South of Germany. If all students within a class worked conscientious, which was indicated by a low percentage of missing data and reasonable time-on-task, the entire class was rewarded a financial support of approximately 150$ paid to the class inventory. Furthermore, students could choose to receive personal feedback on their results. Infrequent cases of missing data due to software problems (e.g., log files were not saved properly) were missing completely at random.

Material

The MicroDYN approach as the measurement framework for CPS, which was administered in this study, represents an innovative approach embedded into LSE systems. The completely computer-based MicroDYN test involves an entire set of independent CPS tasks each lasting approximately 5 minutes. This short time-on-task, which is referred to as multiple complexity in the literature (e.g., Greiff et al., 2012), yields several measurement advantages in comparison to other existing CPS assessments and has proved to be useful in defining the underlying construct on a conceptual and an empirical level

[5] Data obtained from this sample are already used in other publications, mostly to determine psychometrical qualities of the CPS tasks and to cover the between-component perspective. Analyses on task characteristics using CDM as presented in this study are entirely original.

(e.g., Greiff, Wüstenberg, Holt, Goldhammer, & Funke, 2013; Schweizer, Wüstenberg, & Greiff, 2013; Wüstenberg et al., 2012). A comprehensive description of the MicroDYN approach is found in Greiff et al. (2012) and Wüstenberg et al. (2012).

The structure of a MicroDYN task corresponds to the LSE system structure described above (cf. Figure 1). Since MicroDYN is embedded into an LSE framework, it allows the variation of connections between inputs and outputs (i.e. varying number of constituting elements, degree of connectivity, or presence of indirect effects). With regard to semantic embedment, each task has a different cover story (e.g. planting pumpkins, feeding a cat, or driving a moped). To minimize uncontrolled influences of prior knowledge, which is one of the advantages of the MicroDYN approach, inputs and outputs are either labeled without deep semantic meaning (e.g., button A) or fictitiously (e.g., Solurax as name for a fertilizer) as depicted in Figure 2.

The procedure for each MicroDYN task is the same generally applied in LSE systems, in which respondents perform on the two overarching CPS dimensions. In phase 1, knowledge acquisition, respondents first explore the task by manipulating inputs and

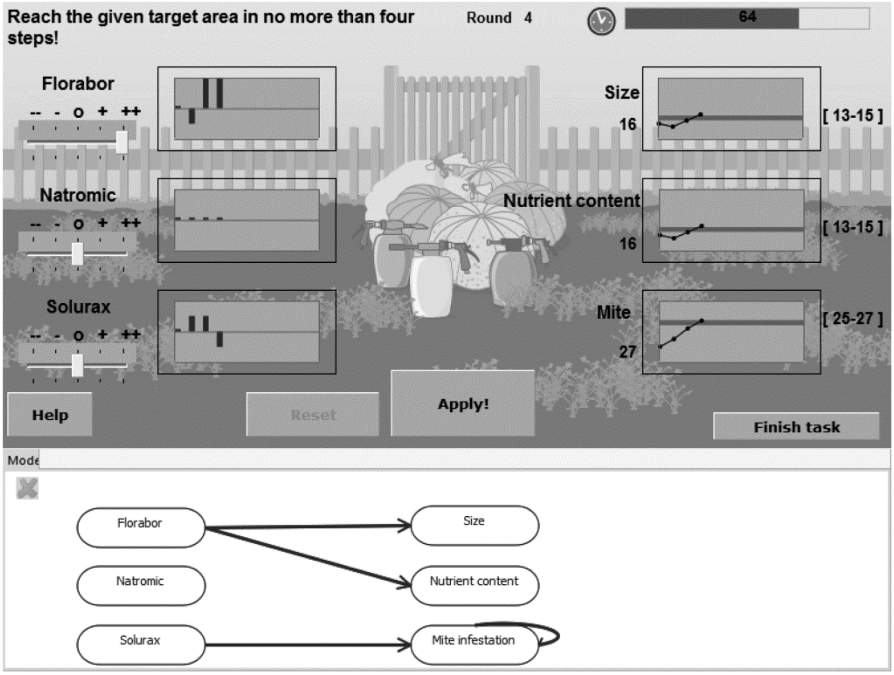
Reach the given target area in no more than four steps!

Figure 2:
Screenshot of the MicroDYN task “Planting Pumpkins” during the knowledge application phase. The participant has to reach the target area indicated numerically and by red areas on the right side of the screen by using knowledge acquired in the first phase and manipulating controller on the left side of the screen.

then represent their knowledge in a causal diagram (Funke, 2001). In phase 2, knowledge application, respondents have to achieve pre-defined target values in the outputs by using the knowledge acquired in the first phase of the task (Funke, 2001; Figure 2). Thereby, in the knowledge application phase, the correct model is depicted as shown in Figure 2, so that the achievement in the second phase does not depend on the correctness of the model drown in the first phase. A MicroDYN test starts with a detailed instruction including trial tasks, in which respondents learn how to operate the software interface, and continues with several independent problems, each of which is administered exactly in the way described above. In the specific MicroDYN test version used, seven MicroDYN tasks were administered lasting approximately 45 minutes.

In order to test the hypotheses postulated above, the degree of connectivity and the presence of indirect effects were varied systematically, whereby tasks included two, three or four connections. For each possible degree of connectivity, except for the very first task with only two effects, there was a task including only direct effects and a task including additional eigendynamics as indirect effects. This way, there were four tasks with no indirect effects and only varying number of direct effects. Three other tasks included indirect effects and varied number of direct effects. Detailed variation of task characteristics for each task can be found in the Appendix.

Procedure and scoring

Test execution was fully computer-based and lasted approximately 45 minutes Participants worked on a set of seven MicroDYN tasks with systematically ex-ante varied task characteristics as described above. In the end, students additionally provided demographical data.

Scoring of each of the seven MicroDYN tasks was categorical, which is an appropriate way to capture CPS performance (Kröner et al., 2005). With regard to knowledge acquisition, full credit was given if models that students provided at the end of the exploration contained no mistakes (i.e., “1”), otherwise no credit was assigned (i.e., “0”). A full score in knowledge application was given if all target values were reached (i.e., “2”), whereas partial credit reflected a tendency to approach but not to fully reach target values (i.e., “1”). If target values were neither reached nor approached, no credit was given (i.e., “0”). For details on scoring consult Greiff et al. (2012) or Wüstenberg et al. (2012).

Statistical analysis and calculation

Besides general descriptive statistics that were calculated for the available data, Rasch models and CDM were applied. For testing the postulated hypotheses, it is a prerequisite that the selected IRT model holds in order to be able to use CDM. This is because in CDM, task difficulty expressed through IRT difficulty estimates is predicted by task characteristics. Thus, before testing CDM, we check for coherence between the Rasch model assumptions and our data. Note that the CDM method used can be applied to the difficulty estimates estimated under any other standard IRT model. We have opted for the Rasch model because of its favorable statistical properties and because MicroDYN is known to be conform with the Rasch model (Wüstenberg et al., 2012).

Of the numerous available CDM, either the LLTM introduced above (Fischer, 1973) or the Least Square Distance Method (LSDM; Dimitrov, 2007) is the natural choice for the present study. Both predict Rasch item difficulties by means of a matrix, in which task characteristics are specified. However, they differ in the assumed link function between task characteristics and difficulty. LLTM postulates an additive link, and LSDM assumes a multiplicative link excluding the possibility of compensation. In a CPS task it is an appropriate assumption that all elements of a problem need to be understood to fully penetrate the problem and that a high understanding of one characteristic cannot compensate a low understanding in another. Thus, for the present analysis the LSDM as a conjunctive multiplicative model was chosen (cf. Equation 4).

$$P_{ij} = \prod_{k=1}^{K} \left[P\left(A_k = 1 \middle| \theta_i \right) \right]^{qjk} \tag{4}$$

Equation 4 represents the probability of correct item response in LSDM – P_{ij} is the probability of correct response on item j for a person at ability level θ_i, $PA_k = \left(1 \middle| \theta_i\right)$ is the probability of correct performance on attribute A_k for a person at the ability level θ_i, and qjk is a 0/1 element in the Q-matrix, allowing all values between 0 and 1 that links item j to attribute A_k (Dimitrov, 2007).

Thus, the LSDM was used to evaluate the impact of specific characteristics of MicroDYN tasks on item difficulty. To this end, the underlying LSE structure (Funke, 2001) of the tasks was varied ex-ante with regard to two characteristics, degree of connectivity (2, 3, or 4 connections) and presence of eigendynamics as indirect effects (not present or present) with all other elements of the underlying structure remaining equal (for specific equations for all tasks see Appendix). Subsequently, the LSDM was used to investigate whether these characteristics affected item difficulty. IRT- as well as all subsequent analyses were conducted separately for knowledge acquisition and knowledge application.

The statistical analysis[6] was conducted with the generalized item response modeling software ConQuest (Wu et al., 1997) and with the R software (R Development Core Team, 2007) package for statistical computing – “Cognitive Diagnosis Modeling,” “Least Squares Distance Method of Cognitive Validation (lsdm)” function.

[6] Acknowledgments for support in statistical issues go to Alexander Robitzsch.

Results

Descriptives

Out of the two MicroDYN dimensions – knowledge acquisition and knowledge application – the knowledge application was easier on a manifest level (*M* = .55; *SD* between tasks = .03; *SD* between persons = .22; α = .68; please note: for this analysis the 0 to 2 range for the scoring of knowledge application was converted to a 0 to 1 range, so that the means between the two dimensions were comparable). Knowledge acquisition showed higher difficulty on a manifest level (*M* = .41; SD between tasks = .08; *SD* between persons = .26; α = .81). Both dimensions showed overall satisfactory α level. Relative frequencies are presented in Table 2 and they indicate that item difficulties differ depending on task characteristics involved in the items. In fact, difficulty in knowledge acquisition and knowledge application varied across tasks and – upon first inspection – was affected by degree of connectivity and presence of indirect effects in items as expected (i.e., eigendynamics; see Hypotheses 1 to 3 for detailed analysis).

Dimensionality of CPS: Rasch model analysis

A Rasch model was applied to the dichotomously scored knowledge acquisition dimension and, separately, a Partial Credit model (Embretson & Reise, 2000) – a direct extension of the Rasch model when more than two categories are used – was applied to the knowledge application dimension, which had three scoring categories. This was performed to check for unidimensionality (i.e., only one latent variable is responsible for performance level in manifest indicators) within each of the two dimensions by forcing factor loadings to be equal, which is only achieved in the context of the Rasch model and its extensions (cf. Kubinger, 2005). Specifically, fit of the Rasch model is a precondition for applying CDM used in our hypotheses, and the Rasch difficulty estimates are further used in the LSDM as dependent variables (Dimitrov, 2007). Since the LSDM uses difficulty estimates as parameters, also the partial credit scoring is allowed, which is the case for the knowledge application dimension (cf. Dimitrov & Atanasov, 2012).

Rasch and Partial Credit model analyses produced item fit indices that were within the endorsed boundaries from .75 to 1.25 (Embretson & Reise, 2000) indicating a good compliance of the data with the assumed models. The Rasch mean square fit values and the item difficulty parameters are presented in Table 2 and they illustrate a satisfactory Rasch model fit for all items. The Rasch correlation between the two dimensions (*r* = .82; *p* < .001) was high. Overall, it was shown that the Rasch assumption of unidimensionality sufficiently held for the items for knowledge acquisition and knowledge application, respectively, a precondition for proceeding with CDM.

Table 2:
Relevant statistics – relative frequency of correct response, Rasch difficulty estimates, and Rasch mean square fit values (MNSQ) – for each of the seven MicroDYN tasks separately for knowledge acquisition and knowledge application.

Task	**Task characteristics**	**Knowledge acquisition**			**Knowledge application**		
		Relative Frequency (1 point) in %	**Rasch difficulty estimates**	**Rasch mean square fit values (MNSQ)**	**Relative Frequency (1/2 points) in %**	**Rasch difficulty estimates**	**Rasch mean square fit values (MNSQ)**
Task 1	2 connections direct effects	51.24	-1.02	1.12	42.09 / 57.26	-2.40	1.00
Task 2	3 connections direct effects	67.28	-2.20	1.12	45.34 / 26.27	0.52	0.94
Task 3	4 connections direct effects	61.35	-1.75	0.85	44.49 / 33.39	0.25	0.88
Task 4	4 connections direct effects	70.31	-2.44	0.98	44.61 / 46.30	-0.73	0.90
Task 5	3 connections direct and indirect effects	5.12	3.70	0.99	40.51 / 34.54	0.20	1.10
Task 6	4 connections direct and indirect effects	18.40	1.54	1.02	53.10 / 4.28	2.04	1.09
Task 7	4 connections direct and indirect effects	13.11	2.17	0.92	37.09 / 39.05	0.11	1.13

Hypotheses 1 to 3

To test Hypotheses 1 through 3, the fit indices for LSDM called Mean Absolute Difference (MAD) were the main criteria. MAD is a nonstandardized value of the mean absolute difference between an Item Characteristic Curve (ICC) and its recovery through LSDM. The ICC recovery through LSDM shows whether the required task characteristics explain the item difficulty across different ability levels, with MAD = 0 indicating a perfect ICC recovery (Dimitrov, 2007). In his study, Dimitrov (2007) offers a classification of ICC recovery as follows: MAD < .02 as a very good recovery, .02 ≤ MAD < .05 indicating a good recovery, .05 ≤ MAD < .10 as a sign of a somewhat good recovery, .10 ≤ MAD < .15 standing for somewhat poor recovery, .15 ≤ MAD < .20 indicating a poor recovery and MAD ≥ .20 indicating a very poor recovery of ICC within LSDM. In order to test whether the recoveries of ICCs through LSDM for both dimensions – knowledge acquisition and knowledge application – were within acceptable boundaries, the mean and median of the MADs across items of each dimension were calculated. Moreover, in order to test how well the two task characteristics combined explain the overall item difficulty, the coefficient of determination R^2 expressing the relation between task characteristics and the item difficulty on an overall level was provided separately for knowledge acquisition and knowledge application.

For knowledge acquisition, the overall fit indices in LSDM, the mean and median MAD, were at .073 and .046, respectively, which is within the commonly applied range of good to acceptable fit (Dimitrov, 2007). Further, both task characteristics exhibited substantial impact on item difficulty ($p < .01$ for degree of connectivity; $p < .001$ for indirect effects). In combination, these two characteristics explained overall item difficulty exceptionally well (coefficient of determination $R^2 = .94$; $p < .001$) with degree of connectivity primarily relevant to differentiate among easy tasks and indirect effects among hard tasks. That is, by increasing the degree of connectivity, difficulty was varied between easy and medium (latent difficulty parameter -1.78 with a mean of 0; DiBello, Stout, & Roussos, 2007), whereas by introducing indirect effects, difficulty was varied between medium and hard (latent difficulty parameter 2.50).

For the other CPS dimension, knowledge application, smaller effects were hypothesized (cf. Hypothesis 3). However, the drop in fit values and effect sizes was surprisingly substantial. Specifically, mean and median MAD were at .145 and .155, respectively, just at the border of poor fit (Dimitrov, 2007). Applying LSDM did not cause any considerable ICC recovery on the overall dimension level. Neither the degree of connectivity (cf. Hypothesis 2.1) nor introducing indirect effects (cf. Hypothesis 2.2) were substantially related to item difficulty ($p > .10$) and the overall determination coefficient R^2 was low at .26 ($p > .05$). This indicates that the two hypothesized task characteristics did not substantially determine the difficulty of knowledge application in CPS tasks.

Summarized, for knowledge acquisition strong effects as expected were observed confirming the hypotheses that degree of connectivity and indirect effects drive the item difficulty of knowledge acquisition (cf. Hypotheses 1.1 and 1.2), whereas no substantial effects were found for knowledge application (cf. Hypotheses 2.1 and 2.2). Thus, Hypotheses 1.1 and 1.2 were supported, whereas Hypotheses 2.1 and 2.2 could not be confirmed. Degree of connectivity and introducing indirect effects drive item difficulty parameters only for one CPS dimension – knowledge acquisition. Hypothesis 3 was also supported, showing even stronger differences in effect sizes between knowledge acquisition and knowledge application than expected.

Discussion

The present study was aimed at examining how ex-ante hypothesized task characteristics determine item difficulties. This served to enhance our understanding of the underlying processes of CPS assessment by using CDM as an advanced statistical method for examining psychometric qualities of the MicroDYN assessment instrument for CPS. The research questions concerned the identification of task characteristics that CPS tasks involved and determining how these task characteristics drove item difficulty parameters of different CPS dimensions – knowledge acquisition and knowledge application. The obtained results provide important information about CPS as construct. They offer guidance on adequate item generation for future research and application endeavors.

While pursuing the within-component perspective of investigating psychometric qualities of CPS, we systematically manipulated two task characteristics ex-ante, which have been proven to be of relevance for mastering CPS tasks in theory and in previous empirical studies – degree of connectivity and indirect effects – in order to investigate how these task characteristics determine item difficulty parameters. CDM revealed that degree of connectivity was generally an easy task characteristic, whereas indirect effects constituted a difficult task characteristic, both being important in mastering CPS tasks. Thus, when learning to deal with complex problems (that is, learning to reduce intrinsic cognitive load imposed by complex tasks; Sweller, 2005), students first acquire cognitive skills allowing them to penetrate increasingly connected systems with direct effects only and later on students start to understand indirect effects (i.e., dynamic developments). Furthermore, according to Bond and Fox (2001) the large difference between latent task parameter estimates for degree of connectivity and indirect effects in LSDM suggests discontinuity of development in CPS (i.e., qualitative changes) albeit standard errors are currently not available for these analyses as statistical backup (Dimitrov, 2007). Overall, effects in LSDM were strong and in line with previous experimental findings for knowledge acquisition, whilst non-significant for knowledge application suggesting other causes for the later CPS dimension. That is, knowledge application performance may reflect implicit learning mechanisms and procedural knowledge, in which demonstration of knowledge may follow other rules than in knowledge acquisition and may not be directly influenced or predicted by task characteristics (e.g., Berry & Broadbent, 1984). Additional possible explanations[7] for the small effects of knowledge application can be found in the potential dependencies between the knowledge acquisition and knowledge application phases. Specifically, in the knowledge application phase, the test taker is provided with the correct model as shown in Figure 2. This model should enable

[7] Acknowlegements to the anonymous reviewer for his/her contribution to this part of the discussion.

the test taker to solve the task even if he/she was not successful in knowledge acquisition. However, we cannot exclude the possibility that test takers who were successful in the first phase versus those who were not may have used different approaches in the second phase. This may lead to the conclusion that other task characteristics (e.g., the complexity of the strategy the test taker uses) may be important for solving the knowledge application part of the task.

Results of this study, suggesting that the item difficulty of knowledge acquisition and knowledge application tasks are driven by different task characteristics of these dimensions, are valuable for future developments in CPS assessment. During the item construction process, other task characteristics for knowledge application need to be taken into account and the task difficulty has to be varied accordingly. Generally, for every measurement tool in psychological assessment, the qualities of the instrument have to be thoroughly investigated in order to provide a high standard assessment. Applying CDM can further enable capturing cognitive processes behind the assessment and to understand how these processes determine the difficulty of the assessment tasks. Furthermore, information about determining task characteristics is valuable for generating new items.

As a word of caution when interpreting results from CDM, it is essential to note that the issue whether results merely reflect task characteristics or allow testimony on underlying cognitive processes in students is highly disputed (e.g., Borsboom & Mellenbergh, 2007). Here is not the venue to resolve this issue, but beyond doubt, the integration of psychometric modeling and cognitive psychology enlightens the substantive understanding of CPS and mastery of its components.

CPS represents a transversal skill, which exponentially gains in importance. A profound understanding of this skill, not only in terms of determining its dimensionality (between-component perspective) but also in terms of discovering the underlying cognitive features of the CPS dimensions (within-component perspective), is essential in order to be able to generate adequate tasks for the assessment of CPS and to enable a meaningful interpretation of assessment results. For instance, the information about involved cognitive processes enables us to explain low results in a CPS assessment as a low ability of a person to perform certain cognitive operations required in the administered tasks. CDM is a promising new field in psychometric research, which offers statistically sophisticated methods facilitating the identification of discrete characteristics that are hiding behind assessment tasks. This way CDM puts a new perspective on the research of CPS, opens up the question of other task characteristics involved in the problem solving process than those handled in this paper, and makes it possible to generate a valid and scalable assessment instrument for CPS.

Author note

This research was funded by grants supported by the Fonds National de la Recherche Luxembourg (ATTRACT „ASKI21“, AFR “COLPASS”). We are grateful to the TBA group at DIPF (http://tba.dipf.de) for providing the authoring tool CBA Item Builder and technical support.

References

Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics, 118*, 1279-1333.

Berry, D., & Broadbent, D. E. (1984). On the relationship between task performance and associated verbalizable knowledge. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 36*, 209-231.

Binkley, M., Erstad, O., Herman, J., Raizen, R., Ripley, M., & Rumble, M. (2012). Defining 21st century skills. In P. Griffin, B. McGaw & E. Care (Ed.), *Assessment and Teaching of 21stCentury Skills* (pp. 17 – 66). Dordrecht: Springer.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Erlbaum.

Borsboom, D., & Mellenbergh, G. J. (2007). Test validity in cognitive assessment. In J. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge, MA: Cambridge University Press.

Buchner, A. (1995). Basic topics and approaches to the study of complex problem solving. In P. A. Frensch & J. Funke (Eds.), *Complex Problem Solving: The European Perspective* (pp. 27–63). Hillsdale, NJ: Erlbaum.

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: examining attributes to a free response listening test. *Language testing, 15*, 119-157.

DiBello, L., Stout, W., & Roussos, L. (2007). Cognitive diagnosis part I: Review of diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 979-1030). Amsterdam: Elsevier.

Dimitrov, D. M. (2007). Least square distance method of cognitive validation and analysis for binary items using item response theory parameters. *Applied Psychological Measurement, 31*, 367-387.

Dimitrov, D. M., & Atanasov, D. V. (2012). Conjuctive and disjunctive extentions of the least squares distance model of cognitive diagnosis. *Educational and Psychological Measurement, 72*, 120-138.

Dörner, D. (1980). Difficulties people have in dealing with complexity. *Simulation & Games, 11*, 87-106.

Dörner, D. (1986). Diagnostik der operativen Intelligenz [Assessment of operative intelligence]. *Diagnostica*, *32*, 290-308.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374.

Funke, J. (1992). Dealing with dynamic systems: Research strategy, diagnostic approach and experimental results. *German Journal of Psychology, 16*, 24-43.

Funke, J. (2001). Dynamic systems as tools for analysing human judgement. *Thinking and Reasoning, 7*, 69-89.

Greiff, S. (2012a). *Individualdiagnostik der Problemlösefähigkeit* [Assessment of problem solving]. Münster: Waxmann.

Greiff, S. (2012b). Assessment and Theory in Complex Problem Solving – A Continuing Contradiction? *Journal of Educational and Developmental Psychology, 2*, 49-56.

Greiff, S., & Funke, J. (2010). Systematische Erforschung komplexer Problemlösefähigkeit anhand minimal komplexer Systeme. *Zeitschrift für Pädagogik, 56* (Beiheft), 216-227.

Greiff, S., Wüstenberg, S., & Funke, J. (2012). Dynamic Problem Solving: A new measurement perspective. *Applied Psychological Measurement, 36*, 189-213.

Greiff, S., Wüstenberg, S., Holt, D. V., Goldhammer, F., & Funke, J. (2013). Computer-based assessment of Complex Problem Solving: concept, implementation, application. *Educational Technology Research and Development, 61*, 407-421.

Greiff, S., Wüstenberg, S., Molnar, G., Fischer, A., Funke, J., & Csapo, B. (2013). Complex Problem Solving in Educational Settings – something beyond g: Concept, Assessment, Measurement Invariance, and Construct Validity. *Journal of Educational Psychology, 105,* 364-379.

Green, K. (1984). Effects of item characteristics on multiple-choice item difficulty. *Educational and Psychological Measurement, 44*, 551-561.

Hartig, J. (2007). Skalierung und Definition von Kompetenzniveaus [Scaling and definition of levels of competency]. In B. Beck & E. Klieme (Eds.), *Sprachliche Kompetenzen. Konzepte und Messung* (p. 83-99). Weinheim: Beltz.

Kluge, A. (2008). Performance assessment with microworlds and their difficulty. *Applied Psychological Measurement, 32*, 156-180.

Kröner, S., Plass, J. L., & Leutner, D. (2005). Intelligence assessment with computer simulations. *Intelligence, 33*, 347-368.

Kubinger, K. D. (2005). Psychological Test Calibration Using the Rasch model – Some Critical Suggestions on Traditional Approaches. *International Journal of Testing, 5*, 377-394.

Mayer, R. E., & Wittrock, M. C. (2006) Problem Solving. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of Educational Psychology* (pp. 287-303). Mahwah, NJ: Lawrence Erlbaum.

Novick, L. R., & Bassok, M. (2005). Problem solving. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 321-349). Cambridge, NY: University Press.

OECD. (2013). PISA 2012 assessment and analytical framework. Paris: OECD Publishing.

Poinstingl, H. (2009). The Linear Logistic Test Model (LLTM) as the methodological foundation of item generating rules for a new verbal reasoning test. *Psychology Science Quarterly, 51*, 123-134.

R Development Core Team (2008). R: A language and environment for statistical computing (Version 2.14.0) [computer program]. R Foundation for Statistical Computing Vienna: Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: a comprehensive review of the current state-of-the-art. *Measurement, 6,* 219-262.

Schweizer, F., Wüstenberg, S., & Greiff, S. (2013). Validity of the MicroDYN approach: Complex problem solving predicts school grades beyond working memory capacity. *Learning and Individual Differences, 24,* 42-52.

Sonnleitner, P., Brunner, M., Greiff, S., Funke, J., Keller, U., Martin, R., Hazotte, C., Mayer, H., & Latour, T. (2012). The Genetics Lab. Acceptance and psychometric characteristics of a computer-based microworld to assess Complex Problem Solving. *Psychological Test and Assessment Modeling, 54*, 54-72.

Sweller, J. (2003). Evolution of human cognitive architecture. *The Psychology of Learning and Motivation, 43*, 215-266.

Sweller, J. (2005). Implications of cognitive load theory for multimedia learning. In R. E. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (pp. 19-30). New York, NY: Cambridge University Press.

Wilhelm, O., & Robitzsch, A. (2009). Have cognitive diagnostic models delivered their goods? Some substantial and methodological concerns. *Measurement, 7*, 53-57.

Wirth, J., & Klieme, E. (2003). Computer-based Assessment of Problem Solving Competence. *Assessment in Education: Principles, Policy & Practice*, *10*, 329-345.

Wu, M. L., Adams, R. J., Wilson, M., & Australian Council for Educational Research. (1997). *ACER ConQuest: Generalised item reponse modelling software manual*. (Version 2.0) [computer program]. Melbourne: ACER Press.

Wüstenberg, S., Greiff, S., & Funke, J. (2012). Complex Problem Solving. More than reasoning? *Intelligence, 40*, 1-14.

Appendix

Seven MicroDYN tasks were varied with regard to two task characteristics: the degree of connectivity between inputs (i.e. A, B, and C) and outputs (i.e., X, Y, and Z with 2, 3 or 4 connections) and presence of indirect effects (i.e., 0 for no eigendynamic or 1 for the presence of eigendynamics).

Task	**Linear structural equations**	**Task characteristics**	
		Degree of connectivity	**Indirect effects**
Task 1	$X_{t+1} = 0{*}A_t + 2{*}\mathbf{B_t} + 1{*}X_t$ $Y_{t+1} = 0{*}A_t + 2{*}\mathbf{B_t} + 1{*}Y_t$	2	0
Task 2	$X_{t+1} = 2{*}\mathbf{A_t} + 2{*}\mathbf{B_t} + 0{*}C_t + 1{*}X_t$ $Y_{t+1} = 0{*}A_t + 0{*}B_t + 2{*}\mathbf{C_t} + 1{*}Y_t$	3	0
Task 3	$X_{t+1} = 2{*}\boldsymbol{A_t} + 0{*}B_t + 0{*}C_t + 1{*}X_t$ $Y_{t+1} = 0{*}A_t + 2{*}\boldsymbol{B_t} + 2{*}\boldsymbol{C_t} + 1{*}Y_t$ $Z_{t+1} = 0{*}A_t + 0{*}B_t + 2{*}\mathbf{C_t} + 1{*}Z_t$	4	0
Task 4	$X_{t+1} = 2{*}\mathbf{A_t} + 2{*}\mathbf{B_t} + 0{*}C_t + 1{*}X_t$ $Y_{t+1} = 0{*}A_t + 2{*}\mathbf{B_t} + 0{*}C_t + 1{*}Y_t$ $Z_{t+1} = 0{*}A_t + 0{*}B_t + 2{*}\mathbf{C_t} + 1{*}Z_t$	4	0
Task 5	$X_{t+1} = 2{*}\mathbf{A_t} + 0{*}B_t + 0{*}C_t + 1.33{*}\mathbf{X_t}$ $Y_{t+1} = 0{*}A_t + 0{*}B_t + 2{*}\mathbf{C_t} + 1{*}Y_t$	3	1
Task 6	$X_{t+1} = 0{*}A_t + 0{*}B_t + 0{*}C_t + 1{*}X_t$ $Y_{t+1} = 2{*}\mathbf{A_t} + 2{*}\mathbf{B_t} + 0{*}C_t + 1.33{*}\mathbf{Y_t}$ $Z_{t+1} = 0{*}A_t + 0{*}B_t + 2{*}\mathbf{C_t} + 1{*}Z_t$	4	1
Task 7	$X_{t+1} = 2{*}\mathbf{A_t} + 0{*}B_t + 0{*}C_t + 1{*}X_t$ $Y_{t+1} = 2{*}\mathbf{A_t} + 0{*}B_t + 0{*}C_t + 1{*}Y_t$ $Z_{t+1} = 0{*}A_t + 0{*}B_t + 2{*}\mathbf{C_t} + 1.33{*}\mathbf{Z_t}$	4	1