Development and Validation of a Problem-Solving Construct

Shafqat Rahman
Patrick Griffin
Zhonghuua Zhang
MGSE, University of Melbourne, Australia

Abstract

This paper explores the concept of problem-solving based on game-like problems not requiring prior knowledge. A set of indicators of student performance generated by Zoanetti (2009) were calibrated using a psychometric model (item response theory). The relationships between the indicators were then explored to determine whether a conceptual model of problem-solving reported by Griffin (2001) could be used to define a construct represented by a developmental progression derived from an item response analysis of indicators in log stream data (Griffin, Mak, & Wu, 2006). A latent trait analysis approach was used to calibrate the indicators and to determine properties of the construct in terms of reliability and validity of derived quality measures. It defined a learning continuum of cognitive processes that could be used to assist teachers to determine students' zone of proximal development (ZPD) to measure progression and opportunities for effective intervention. The initial analysis confirmed the hypothesis of the study. This study established a methodology to develop and validate interactive problem-solving progression. It further validated that this hierarchal model best fits the indicators extracted from the log stream file within a psychometric model.

Introduction

In the educational field, problem-solving is considered an essential skill for an individual to be able to adapt to a rapidly changing society. Societal needs and challenges in the real world demand changes in the school curriculum and simultaneously, require appropriate processes and procedures to be adopted for the development of cognitive skills. The curriculum must widen its scope by including thinking strategies, collaborative ways of working, methods for the application of modern technology, and productive ways of living in the world. It is internationally accepted that there is a need for the revision and redesign of the curricula to articulate the fundamentals of 21st century skills (Griffin, McGaw, & Care, 2012) and to include models and appropriate learning activities (Partnership for 21st Century Skills, 2009) to prepare students for a technologically growing world. These models and activities should enable students to think critically and act logically to evaluate situations (O'Neil, 1999), solve problems, and make decisions, allowing them to succeed in their future workplaces.

Background of the Study

A review of research in the field of problem-solving has indicated that studies have uncovered only pieces of the puzzle concerning the construct of problem-solving , suggesting that further research should be examined together to obtain a complete picture. Most problem-solving studies are dominated by mathematics and especially by the mathematical problem-solving model popularised by Pólya (1973) and Schoenfeld (1985), who adopted mainly domain-specific approaches. Pólya (1973) proposed a sequential phase approach towards problem-solving , which today remains an assessment guide in educational research.

Chi, Feltovich, and Glaser's (1981) work focused specifically on curriculum-based problem-solving in the area of physics, whereas, O'Neil's (1999, 2002, 2003) focused on computer-based performance assessment and especially information-seeking strategies. Csapó's (2007) research examined students'

cognitive skills, adopting a Piagetian or neo-Piagetian approach to problems solving. However, Funke (2010) assessed dynamic problem-solving (also called complex problem-solving) through psychological tests (MicroDYN) with problem-solving performance items used in experimental research (Funke, 2001). MicroDYN was also adopted for the Programme for International Student Assessment's (PISA) 2012 problem-solving assessment (Organisation for Economic Co-operation and Development [OECD], 2013).

While both Wu (2003) and Zoanetti (2009) reviewed the cognitive processes in problem-solving , Wu (2003) applied them to the specific domain of mathematics by applying paper-and-pencil testing using Pólya's (1973) sequential process-based approach. She introduced psychometrics to the assessment of problem-solving by fitting a psychometric model to coded student performances. Similarly, the PISA project (OECD, 2003, 2007, 2013) assessed problem-solving using a conceptual framework based on Pólya's (1973) four-step mathematical problem-solving model using mathematical or pseudomathematical tasks. The Assessment and Teaching of 21st Century Skills (ATC21s) project (ARC, 2009–2012)identified individual problem-solving as an area to be explored, but due to time constraints the idea was abandoned in 2010. The success of the ALP project, which was initiated by the ARC and the Catholic Education Office, Melbourne (ARC, 2004), has led educational research towards developmental learning paradigms. The project's findings showed that teachers who apply developmental learning frameworks using assessment data increased their influence on student learning achievements (Griffin et al., 2013). Three developmental approaches—Bloom (1956) for knowledge; Krathwohl, Bloom, and Masia (1964) for attitude development; and Dreyfus and Dreyfus (1980) for skill development—were used in parallel with Vygotsky's approach (1978) to identify a student's zone of proximal development (ZPD) within the knowledge, skills, and attitudes frameworks. Vygotsky's (1978) ZPD theory and Glaser's (1963) theoretical framework, which interpreted students' levels of increasing competence within a Rasch measurement model (Rasch, 1960), were combined by Griffin (2007a) to create a developmental model that provided new opportunities for educational researchers to assess students' learning.

It can be argued that promoting individuals' problem-solving skills is an important element in supporting the development of collaborative problem-solving. In Zoanetti's (2009) study, problem-solving was assessed using a series of game-like puzzles, which moved problem-solving away from curriculum dependence and from the Pólya (1973) approach, which was fundamentally based on mathematical problem-solving (explored extensively by Wu, 2003). Zoanetti (2009) presented key steps in design and analysis of computer-based problem-solving assessment using Bayesian inference networks. However, Zoanetti (2009) did not focus on identifying the construct underpinning student performance on these problem-solving tasks, which could be fitted to the psychometric model instead of, or in addition to, a Bayesian network, which is purely statistical in nature. The tasks used by Zoanetti focused on hypothetico-deductive thinking (Griffin, 2014); therefore, this study adopted these tasks to explore the underpinning construct and to observe the growth in inductive and deductive reasoning skills.The developmental approach and Zoanetti's games like interactive tasks have provided a foundation for the current study in the field of domain-general problem-solving and, specifically, in complex, interactive problem-solving to scaffold learning.

Literature Review

The ability to solve problems is general and modifiable (Adey, Csapó, Demetriou, Hautamäki, & Shayer, 2007) and provides a number of opportunities in the field of education, noticeably contributing to knowledge and its application in new situations (Csapó, 2007). Over the last two decades, the need to improve students' problem-solving skills has been greatly emphasised. In the last decade, there have been two trends in the theories of problem-solving : domain specific and domain general. However, there has recently been a shift from domain-specific conceptualisations of problem-solving towards domain-general models (Jonassen, 2000).

Problem-solving skills improve performance in both specific learning areas such as mathematics, and

in general learning contexts (Zoanetti, 2009). In order to avoid the limitations of both approaches, it is recommended (Tricot & Sweller, 2014) that students should learn in diverse situations where they can apply an already learned generic skill. It is then preferable to focus on learning a generic skill, rather than focusing on domains. This could be applied to solving problems in domain-specific situations as well as in everyday life. Furthermore, Limjap (2011) suggested that problem-solving research should focus on learning processes and how they relate to overall problem-solving . Therefore, a shift from a domain-specific to a domain-general approach in problem-solving is underway. Thus, Zoanetti (2009), Funke (2001) and Crisp (2010) introduced and developed interactivity by using "physiognomies" in the new era of the computer environment in education, by developing interactive problem-solving assessment tasks. This kind of educational setting provided the opportunity for students to establish and improve skills to evaluate their own competency levels (Boud & Falchikov, 2007), which is critically important for enhancing efficient methodologies for further learning.

Problem-solving as a construct

Many international organisations and projects are searching for valid indicators of student achievements and setting new goals for education in the field of problem-solving. Problem-solving is a natural part of everyday life. Irrespective of field, a problem is defined as something that occurs and there is no feasible or routine way to resolve it (Greiff, Holt, & Funke, 2013). Problem-solving involves decision-making skills and techniques that are often closely linked to how an individual opts to solve the problem (Greiff et al., 2013; Ohlsson, 2012).

The PISA project (OECD, 2013) showed that problem-solving in a collaborative format has distinctive advantages where it allows for a coordinated construct and shared understanding of the processes needed to resolve it. Problem-solving is a productive ability to demonstrate cooperation and the ability to work in groups to organise steps in resolving issues. Problem-solving is an assessment of a situation and subsequently, it requires elements to be broken down for effective resolution, between both the individual and the group (OECD, 2013).

Although these definitions emphasise the need for a goal state and the application of cognitive processes in a problem-solving activity, the actual nature of the cognitive processes and skills necessary to achieve the goal are missing. Even domain-specific, diagnostic problem-solving may possibly encompass multifaceted overall cognitive processes that are not restricted to specific domains (Funke, 2001; Sternberg, 1996). The PISA 2012 (OECD, 2013) field trial of computer-based interactive problem-solving assessment, in contrast, aimed to develop an understanding of and measure individuals' problem-solving competencies in general. However, most problem-solving studies focus on the same construct that restricts the model to the Pólya (1973) sequential step construct.

Conceptual Model

In this study, the construct and conceptual model was hypothesised to be hierarchical and developmental in nature. This developmental approach was initially proposed by Griffin (2001, 2014) and investigated by Callingham (2004), who examined the competence of students in complex problem-solving numeracy tasks using this approach. Although the higher-order thinking skills construct adopted in Callingham's (2004) study was limited to the numeracy context, it is considered relevant for this study because of the idea that this construct is inbuilt in the general problem-solving game-based tasks developed by Zoanetti (2009). Therefore, this study's literature review focused on the developmental continua of competence.

Developmental frameworks

Developmental frameworks could be recognised as developmental taxonomies, such as: (1) Bloom's revised taxonomy (1956), Dryfus's model of skill acquisition (1980), structure of the observed learning outcomes (SOLO) taxonomy (Biggs & Collis, 1982, 1991); (2) hypothetical progression, which is an adaptation of any of the developmental taxonomies, particularly in the context of learning areas; (3) curriculum progressions based on curriculum standards such as AusVELS(2014); (4) derived progression such as NAPLAN, Assessment Research Centre Online Testing System numeracy, reading and creative problem-solving, collaborative problem-solving 21st century, and Students With Additional Needs (SWAN) progression (Hutchinson, Francis, & Griffin, 2014); and (5) the continuum of competency in numeracy contexts (Callingham, 2004; Griffin, 2000, 2001, 2014). All the taxonomies have many advantages; however, in order to link students' scores with their ZPD, it is important to adopt a model that has the potential to identify students' skill levels and ZPD. Consequently, it is essential to explore Glaser's (1963) stages along the progression of increasing competence and Griffin's continuum of competence.

Glaser's model

Glaser (1963) was the first to introduce the criterion-referenced framework. He interpreted the assessment data in relation to the inherent potential and organisation of the tasks accomplished. He described the skills embedded in the task to be established by the student. Therefore, students' performance and development was interpreted according to this criterion, and hence termed the "criterion-referenced framework."

Griffin's hypothetico-deductive continuum

Griffin (2007a) proposed another theory of learning by linking Vygotsky's (1978) constructivist theory of ZPD with the probability of success for each person, which is 0.50. He linked Rasch's (1960) measurement theory with Vygotsky's (1978) ZPD, where success had a probability of .5; this indicated the person's location on a continuum of latent trait, providing the best chance for teaching intervention. Glaser's (1963) stages along a progression of increasing competence and its criterion-referenced interpretation were also combined (Griffin, 2007a) with Vygotsky's ZPD, and these were further linked to latent trait theory by matching the ability of the students with the difficulty of the task (Hutchinson et al., 2014). Griffin's (2007a) work has great implications in the field of educational assessment and learning, where the notion that "assessment is for learning" was transformed into "assessment is for teaching".

Griffin (2009) highlighted that the combination of the latent trait theory and criterion-referenced test interpretation offers an opportunity to connect a person's skill level to a scale of developmental competence. The scale pinpoints skills progression and an individual's location on the scale. This establishes meaning to the score as well as the underlying construct by interpreting capabilities needed to respond correctly to the item. Therefore, Griffin (2014) argued that the process dimension could be simplified with fewer levels. His proposed framework emphasised that problem-solving can be seen as a hierarchy of different levels of proficiency, moving from inductive at the lowest level to deductive reasoning at the highest level. This is also reflected in the descriptions of problem-solving in the ATC21s project, where it is defined as a process where the

> problem solver first examines the problem space to identify elements of the space. Next they recognise patterns and relationships between the elements, and formulate these into rules. The rules are then generalised and when generalisations are tested for alternative outcomes the problem solver is said to be testing hypotheses. (Hesse, Care, Buder, Sassenberg, & Griffin (2015) p. 4)

The simplified hierarchy is illustrated in Figure 1.

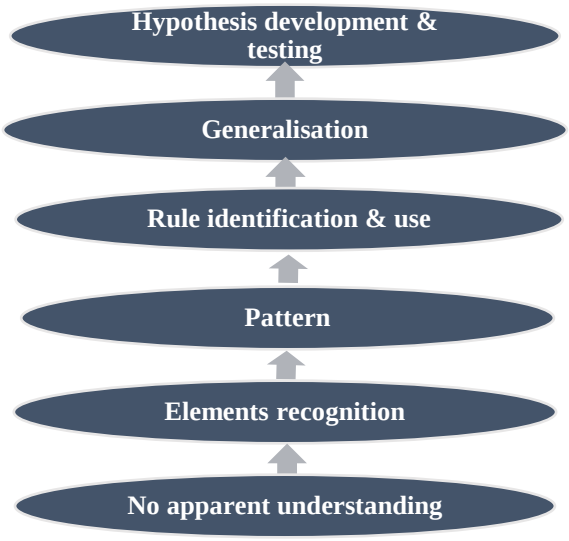

Figure 1: Hypothetical Continuum of competence described by Callingham and Griffin (2000, 2001) and Griffin (2014).

The conceptual definition of the underlying latent construct adopted in this study is the empirical continuum of competence described by Callingham and Griffin (2000, 2001) as the manifestation of a latent variable, as detailed in Table 1. This leads to the development of the hierarchical construct of elements -› pattern -› rule -› generalisation -› hypothesis that characterises the continuum of competence.

Up until recently, several researchers used a number of frameworks to measure a student's problem-solving ability. However, the current research is an attempt to shift the approach towards generalised sequences of cognitive development that employ higher-order hypothetico-deductive thinking skills as described by Griffin (2014). It is likely that the construct explored by Callingham (2004) in a numeracy context could also be observed in problem-solving skills in the general domain. Moreover, problem-solving could be seen as a single construct that can be inferred from the coded behaviours exhibited by the students in problem-solving interactive tasks such as those developed by Zoanetti (2009).

Table 1. Empirical Continuum of Competence (Callingham & Griffin, 2000, 2001)

Making conjectures At this level the student can suggest extensions of the original problem, or change the task parameters to create new situations by posing 'What if' questions. Problem solutions are expressed in appropriate symbolic or technical language. The student is ready to learn how to identify assumptions and develop further hypotheses.
Relationship or generalisation use At this level the student can express the generalisation of the problem solution in symbolic form, and apply it to new situations, justifying this by reference to the generalisation. The student has mastered the particular problem type and is ready to learn how to change the problem type to explore new ideas.
Relationship or generalisation recognition At this level the student can form a generalisation of the solution strategy and express this generalisation in words. The student is ready to learn the use of symbolic forms and technical language that will allow transfer of the generalisation to other settings.
Rule or process use At this level the student's own rule is applied to extensions of the initial task. The student can extend the solution obtained to a limited range of other tasks having a similar structure to the initial one, and is ready to learn how to form a generalisation that could be transferred to other settings.
Rule or process recognition At this level the student recognises a rule underpinning the structure of the task and is ready to learn how to apply that rule consistently and extend the use of the rule. The rule is likely to be expressed in words or

diagrams, using non-technical language that summarises the student's own approach to the problem.
Pattern or structure use At this level the student recognises the underlying principles in the structure of the task, and can apply these in a familiar setting, such as a straightforward extension of the initial task. The student is ready to learn how to identify a rule that links the repeating elements together.
Pattern or structure recognition At this level the student recognises the repeating elements in the structure of the problem, and is ready to learn how to recognise the underlying principles. Problem solutions are likely to be presented as incomplete diagrams or oral explanations.
Element recognition At this level the student recognises individual elements of the task and is ready to learn how to combine these into a pattern or structure. Problem attempts are presented as single drawings or phrases that relate to one element only of the task.
No apparent understanding At this level there is not enough information to describe the student's work. If the task was attempted, it is likely that the student did not recognise the elements of the underlying structure of the task.

There is little exploration in the literature of the types of task that could prove that problem-solving is a set of cognitive processes moving from inductive to deductive reasoning. Therefore, this research attempted to develop a learning continuum of cognitive processes to assist teachers to determine students' ZPDs (Vygotsky, 1978) and to measure their progression for effective intervention. This study aimed to identify the construct of problem-solving based on two hypotheses:

1. A psychometric model (Rasch, 1960), which could be used to interpret the data (Griffin et al., 2006) obtained from problem-solving game-based tasks.

2. A conceptual model (Griffin & Callingham, 2000, 2001; Griffin, 2014), which could be used to fit the indicators extracted from the log stream file in the psychometric model.

Indicators and Cognitive Processes

The starting point for this research was to collect information from the available literature as well as to study student responses from an existing problem-solving data set (Griffin et al., 2006). In order to identify the evidence of the construct and its measure, and to use that to improve students' problem-solving skills, it was essential to explore the processes and skills that can be transformed directly into instructional strategies. Importance was placed on activities such as random and systematic trial and error, exploration, observation and discovery, pattern recognition, rule formation and making, and generalising of hypotheses.

Empirical information about the target variables as indicators in this study was identified in a literature review by Zoanetti (2009). These target variables are further illuminated with relevant reference to research concerning problem-solving strategies and cognitive processes. Initially, Zoanetti (2009) categorised these indicators into the profile variables based on explanations from theory and practice: decoding time, initial representation, error tendency, attainment, activity, and search duration. The profile variables were further separated into observable variables (Table 2), which were then used in the current study as indicators of students' behaviour.

Table 2. Profile Variables and Observable Variables Derived from Zoanetti (2009)

Profile Variables	Observable Variables	
Decoding time	***Pre-search latency (PSL):*** Time taken before starting to solve the problem	
Initial Representation	***Best-first Action (BFA)***	***Valid-first Action (VFS)***
	The first action or set of actions is considered valid if it complies with task instructions.	
Error Tendency	***Invalid action count (IAC):*** Invalid actions refer to those moves, or actions, that violate the constraints or against the rule of the game.	

Attainment	***Goal attained/ Correctness:*** To submit the solution after achieving the goal.			
Activity	***Search scope (SS):*** Search scope refers to the number of distinct states visited in data. In other words, the search scope can be described as possible steps or actions that the problem solver might take to reach the goal.	***Action count/Number of actions (AC):*** The count of clickstream interactions with task. This is extracted from the log stream file by subtracting the total rollover count from the total actions taken.	***Repeated action count (RAC):***	***Rollover count (RoC):*** The number of times the mouse is rolled over the submit button.
Search duration	**Response Time (RT):** Search duration to finish the task			

Research Questions

Two identifiable issues remain after the work of Zoanetti (2009): a lack of developmental progression as a manifestation of the latent construct, and the shift from four-step mathematics tasks to game-like problems. These gaps led to the following research questions:

RQ1. What is the nature of the construct of problem-solving when game-like tasks are used?
RQ2. To what extent can meaningful and reliable indicators be developed to measure problem-solving abilities using indicators from log stream data generated by students completing game-like tasks?
RQ3. To what extent can the hypothetico-deductive continuum be applied to the data generated from game-based problem-solving ?

Methodology

The research questions addressed the overall definition of problem-solving when students were assessed using game-like tasks. They were considered as empirical in order to investigate the theoretical continuum of competence by mapping secondary data onto a psychometric model to draw a progression. Therefore, the study employed a criterion-referenced developmental progression approach and used item response modelling (Rasch, 1960) to achieve that end. The observations in this case were initially linked to those students who attempted to solve game-like problems. The data used in forming the developmental progression are the data from project initiated by Griffin et al. (2006) and collected by Zoanetti (2009). The data are used for the initial development of the framework and to obtain evidence of its empirical construct. The data are modelled using Rasch (1960) to identify an underlying developmental continuum representing a problem-solving construct. The Rasch (1960) model is used to estimate students' abilities and item difficulties, and to draw inferences about students and item characteristics. Furthermore, these inferences can be linked to educational measures, such as actions to intervene, in the learning process of students.

The methodology of this study follows a logical sequence of six stages divided into two phases. In the first phase, it describes briefly the data exploration, coding, and scoring procedure, followed by the calibration and interpretation of the variables to map the indicators onto the measurement model and conceptual model. In the second phase, it deals with the validation of the outcomes and results. This paper discusses the first phase of the methodology that deals with the development of the problem-solving progression.

Game-based problem-solving tasks

The game like tasks were designed using Macromedia Flash 8 and embedded within an HTML web page. These tasks were developed by emphasising procedural knowledge over declarative knowledge (Zoanetti, 2009). The design of the tasks required that each problem should be a multistep problem rather than single-step. These were interactive, well-defined problem tasks with a clearly specified goal to capture process data that can only be recorded in a computerised situation. The tasks (Figure 2) were constructed in such a way that the objects could be dragged and dropped by using the mouse curser, without using the keyboard.

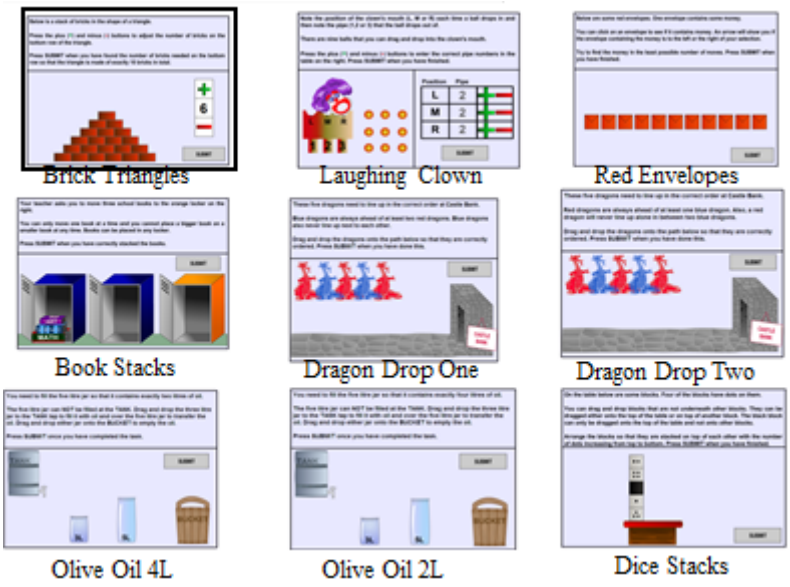

Figure 2. The interface of game-based problem-solving tasks

Problem-solving tasks in this study were designed in a way that students must intermingle with the objects in the tasks to solve them. Interactions with the task were recorded using ActionScript 2.0 syntax. For instance, Book Stacks tasks in Figure 3 required the problem solver to rearrange the configuration of task objects (books) in one dimension to match explicit criterion in the task instructions (Zoanetti, 2009).

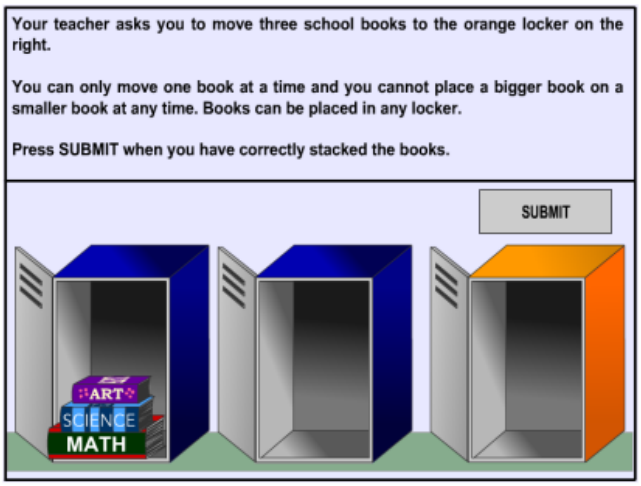

Figure 3. Book Stacks task from Task Model 3. Source: Zoanetti (2009, Appendix II)

Data Description/Exploration

This section investigates the process of data exploration to analyse the log stream data (Tables 3 and 4). The data file (Zoanetti, 2009) contained log stream files, that is, click stream/process data comprised distinct mouse clicks, drags and drops, rollover, cursor movement, and so on, where each discrete action was recorded with a timestamp. The timestamp refers to the time when a click event was recorded by the system into a log file.

Tables 3 and 4, the "work product," indicate that the student attempted task 0099 (which is the Olive Oil 4 Litre task). The student submitted a response of 5 litres, which is the wrong solution. The student spent 163.071 seconds engaged in their solution attempt and, as an example of one particular interaction, after 56 seconds the student emptied the 3 litre jug (56_e5) a second time (Zoanetti, 2009).

Table 3. Example of Log Stream Data File

Student ID	Task ID	Task	Date and time data collected	Student response/log stream file
Unknown	0099	Olive Oil 4L	24-Oct-2008 02:45:00	it_0099:res_5:rt_163.071:sq_!21_f3!28_t3!36_e5!47_roSub!56_e5!56_e5!61_f3!85_f3!90_t3!94_e5!107_f3!113_t3!122_f3!127_t3!132_e3!142_t3!147_f3!15_t3!151_t3!161_e3!::!!

Table 4. Example of Log Stream Data and Explanations for Actions Recorded

Actions/log stream data	Sequence of actions /description	Quality of actions/indicators
res_5:rt	Response submitted is 5 (incorrect solution)	Correctness
21_f3	Fill 3L jug in 21 seconds	Valid first action
28_t3	Transfer 3L in 28 seconds	Best first action
36_e5!	Emptying the jug in 36 sec	Valid action
47_roSub	Rolling over mouse	Thinking/distracted
56_e5	Emptying the jug	Invalid action
61_f3	Filling 3L jug again for second time	Valid
85_f3	Filling 3L jug again for third time without emptying	Invalid action
90_t3	Transferring 3L jug	Repetition of actions
94_e5	Emptying 5L jug	Random trial and error
107_f3	Filling 3L jug for fourth time	Repetition of actions
113_t3	Transferring 3L jug	Random trial and error
122_f3	Filling 3L jug for fifth time	Repetition of actions
127_t3	Transferring 3L jug	Random trial and error
132_e3!	Emptying 3L jug	Random trial and error
142_t3	Transferring 3L jug	Random trial and error
147_f3	Filling 3L jug for sixth time	Repetition of actions
151_t3	Transferring 3L	Random trial and error
161_e3	Emptying 3L jug	Random trial and error

Various response-processing algorithms (e.g., Figure 4) were developed by Zoanetti (2009) to facilitate the conversion of work products into values for observable variables.

```
/*Item ID*/
                    var item:String="0057";
```

```
            /*response product*/
            var res:String=nlitres;
            /*response time*/
            var rt:Number;
            rt=0.001*getTimer()-starttime;
            /*response sequence*/
            var sq:String=sequence;
/*This function describes how the data will be structured prior to sending to PHP*/
function Package(item:String,resp:String,rt:Number,seq:String) {
varsToSend = new LoadVars();
varsToSend.Response="it_" + item + ":res_" + resp + ":rt_" + rt + ":sq_" + seq;
varsToSend.send("ItemProc.php","_self","POST");
}
Package(item,res,rt,sq);
```

Figure 4. Response processing was handled using scripts within tasks and on the server (Zoanetti, 2009, p. 39).

The observable variables captured from the process data are labelled here as "indicators" or "observable behaviour." For instance, Tables 3 and 4 log stream actions provide some explanation for the type of data and student's behaviour.

Coding

The point of departure in this study was the focus on how students solve problems rather than only recording and coding their success or failure in solving problems (Adams, Vista, Scoular, Awwal, Griffin, & Care, 2015). Indicators extracted from the log stream file were coded and scored to construct a distribution of values.

For example, the Laughing Clown task (Figure 5) was coded *LWC* and any indicator of actions taken to explore the problem space by the students was coded *AC*, resulting in the code *LWC_AC*. Variables were systematically named for ease of identification and to simplify the nomenclature convention. Although the same original nomenclature as used by Zoanetti (2009) was used for the indicators, only the codes were changed for the purpose of simplicity.

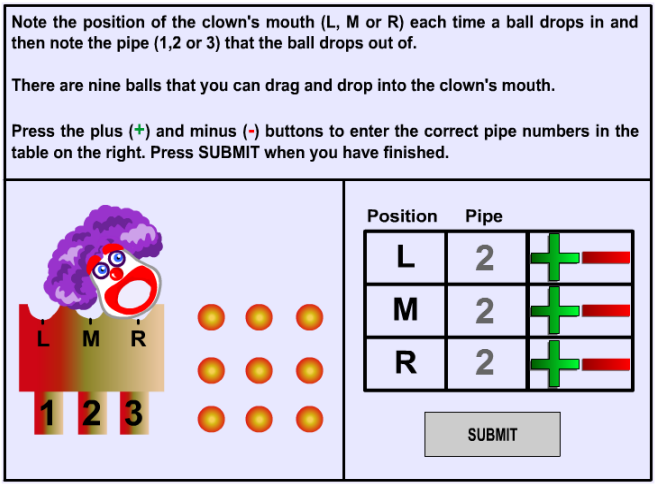

Figure 5. Laughing Clown task.

Scoring Rule

Sequences of actions led to inferences, which signified specific skills and levels as well as direct

actions. In order to link instructional decisions to evidence of process, the recorded data should have a system of decoding in terms of skills, termed as rubrics, for assessment purposes; a rubric refers to the "scoring rules" (Gillis & Griffin, 2004). In this study, the scoring rules referred to the performance rubrics or criterion-referenced measures, whereby a performance rubric is a key component in many authentic assessments.

There were two types of data: dichotomously coded/scored (0, 1), and other carrying thresholds values. However, indicators with distribution properties were converted into scoreable levels of performance using raw data. The resulting distributions were used to score each indicator into qualitative levels in order to interpret those levels or indicators as ability scores. This meant that students were given a score of 0, 1, or more for each action, depending on the amount of proficiency displayed. For example, students could receive a score of 0, 1, or 2 for any indicator with threshold values, with each of these scores indicating that they had demonstrated progressively more competent levels of proficiency (Griffin et al., 2014). For example, the threshold values of pre-interaction times were initially scored as ordered polytomous scores, that is, $\delta_3 > \delta_2 > \delta_1 > \delta_0$ into four levels (e.g., 0–10s = 0, 10.1–40s = 3, 40.1–90s = 2, > 90s = 1).

An histogram of pre-interaction time (Figure 6) for the Laughing Clown task shows three points on the Figure 4, of X, Y and Z. Therefore, threshold values were scored as:

IF PSL ≤ X seconds then less time/very low =0
IF X ≤ PSL ≤ Y then optimum time/high =3
IF Y ≤ PSL ≤ Z then appropriate time/intermediate=2
IF PSL > Z THEN extended time/low =1

where in this case X = 10 sec, Y = 40sec, and Z = 90sec.

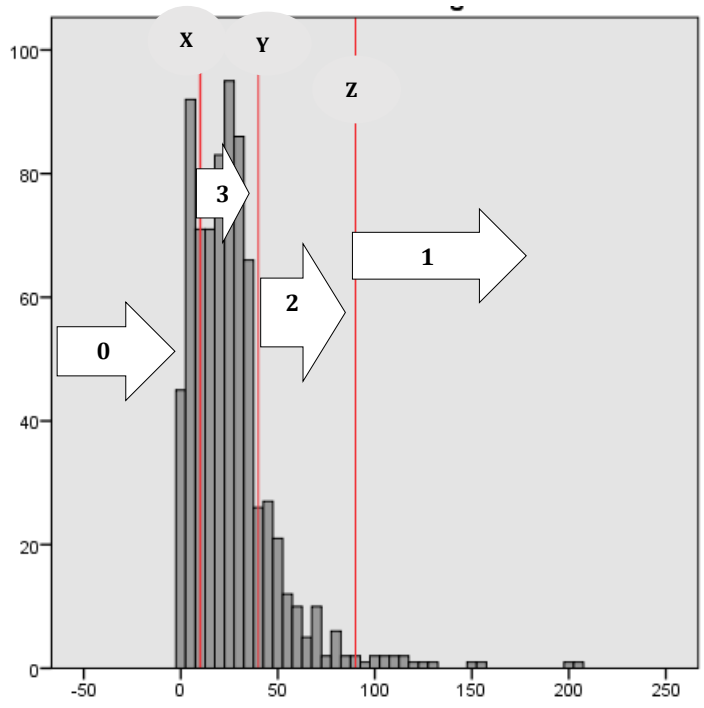

Figure 6. Sample histogram of a pre-interaction time.

Consequently, it is evident that instead of scoring the responses, students' interactions were coded for certain skill levels rather than for their scores only (Table 5). Each code set that was provided as a way of recording the different levels of quality were evident in students' performances. Each coding scheme was internal to the indicator within the task such that a code of 1 referred to the first level of

performance on that indicator and did not relate to the quality or value intended for a code of 1 for any other indicator.

The greatest challenge in observing and judging complex behaviours is to link them with reliability and validity (Griffin & Robertson, 2014), for instance, students' interactions in the current study. This study therefore used advanced psychometric methods throughout the scale development process to provide a clear understanding of the latent construct, particularly with respect to the populations studied, and to develop adaptive and nonadaptive instruments with appropriate psychometric properties for implementation in a range of research applications. This process comprised the analysis of items to check the reliability of scores and properties of the problem-solving scale using psychometric methods e.g., item discrimination, point-biserial, item difficulty parameters (delta) and item fit (statistics). The scores were finally changed to three levels 0, 1, and 2 after deep analysis of the psychometric properties of the items.

Table 5. Explanation of Indicator Codes and Levels with Scores

Indicators Name	Code	Explanation	Level 1 Score (0)	Level 2 Score (1)	Level 3 Score (2)
			Rubrics		
Actions Count	**AC**	The number of activity/or actions taken indicates their efficiency in exploring and reaching the goal. However, to solve problem requires a full range of possible exploration of the problem space.	Redundant interactions/no efficiency/no evidence of exploration	Moderately efficient/random trial and error	Highly efficient/systematic trial and error
Valid first action	**VF**	Forward working and goal directed and goal oriented	Unplanned/goal lost	Goal directed/planned	
Best First action	**BF**	Problem solver is following the task instruction carefully, evident from their very first activity.	Unplanned/no understanding of the task	Planned/clear understanding of the task	
Correctness	**Corr**	Correct solution	Incorrect	Correct	
Time taken before interacting with the task	**PSL**	Time taken to decode or understand the problem shows that their efforts begin with a speculation in planning and analysis.	No investment in planning	Limited efforts and time to understand the problem	Optimum efforts and time/deep understanding of the problem
Response Time	**RT**	Response time in combination with correctness or number of actions could be used as an indicator of expertise and efficiency. This time along with before interaction time indicates the problem representation with appropriate planning and execution of goal.	Time taken to submit the task is very low and limited not enough for exploration. Or, in another case some students are confused and taking an extended amount of time, trying to solve problem with trial and error.	Time is intermediately enough to explore. Some random trials and evidence of planned attitude.	At this level, students will spent optimum time that indicates this group of students has taken optimum required time to plan and execute the goal appropriately.

Rollover counts	**RoC**	The students who are confident about the correctness of their solution show no rollover or less rollover attitude as they have now started making to make a judgement about the solution path.	Confused/ distracted/ uncertainty Guess and check	Focused to some extent/ some distraction/ minimal uncertainty	Completely focused/confident with no uncertainty about their solution strategy.
Invalid Actions Counts	**INC**	Expert problem solver is able to manage errors and avoid actions that are against constraints.	Flawed interactions/unable to manage errors.	Able to some extent to avoid errors and show error management.	Flawless interactions/ high ability to manage and avoid invalid actions that is against the constraints.
Search Scope	**SS**	Indicates the trial to test all possible solutions in order to recognise the solution pattern. This also indicates systematic trial and error approach.	No evidence of testing solution.	Some of the solution path is tested. Random exploration or trials.	All possible solution tested. Systematic exploration and testing solution path.

Calibration

Under the assumption that all indicators measured the same construct, the data were jointly calibrated on a latent uniform proficiency scale by fitting a unidimensional partial credit model (PCM) (Masters, 1982) to the students' responses of 60 indicators using ConQuest (Wu, Adams, & Wilson, 1998; Wu & Adams, 2007). ConQuest produced tables for item step difficulty (logit), measurement error (SE) for each step difficulty, the weighted fit (infit) estimates and the confidence intervals of these estimates along with the t value.

Thus, in the context of this study, fit indices were examined to investigate the degree of correspondence between the conceptual model and the measurement model predicted by the Rasch model. The mean weighted fit (infit) of the final data set was 0.999 with a variance of 0.0003. The item fit statistics, (infit MNSQ) ranged from .96 - 1.06 (0.7–1.3), indicating that the data fit the model's expectation very well. It confirmed that a single construct being measured provided evidence of construct validity (Wolfe & Smith, 2007).

ConQuest (Wu et al., 1998) also produced a variable map (Figure 7), which in turn provided the relationship between tasks and persons (Griffin, 2007a). The variables map indicated the relative positions of the students' abilities relevant to problem-solving. Starting from the left of Figure 5, the first distinguishing element of the chart is a scale that ranges from –4.0 to +4.0. This is the logit scale and is the metric of the Rasch analysis that empowers both item difficulty and student ability to be mapped onto the equivalent scale.

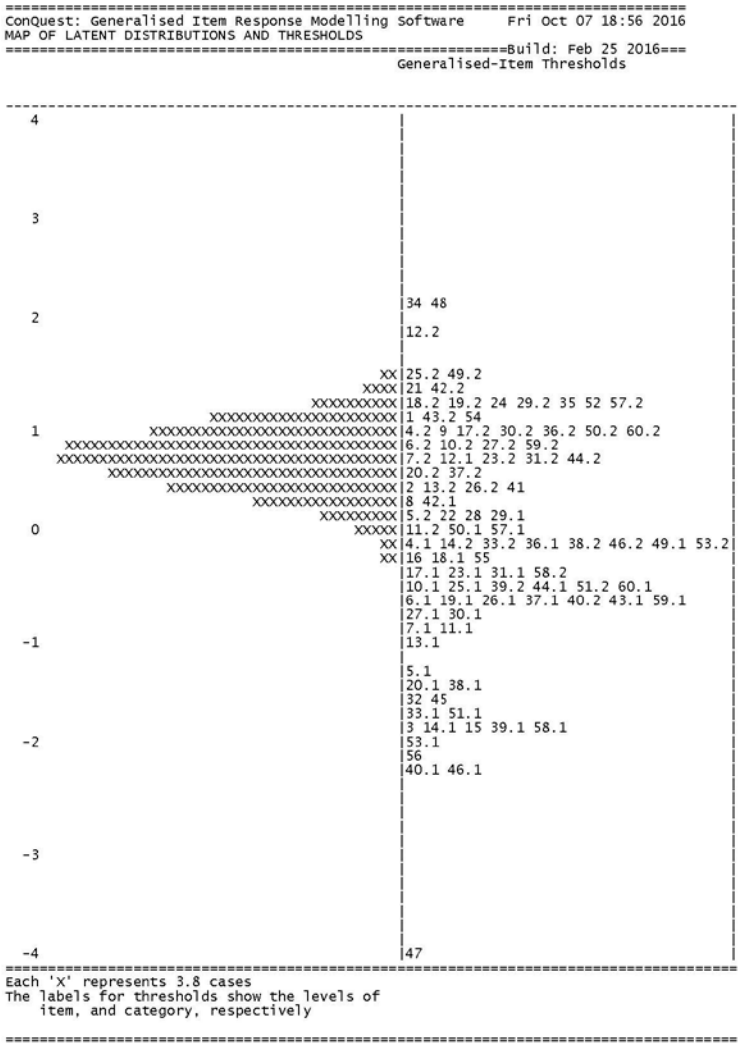

Figure 7. Variable map for a set of nine tasks as assessment instrument.

A person separation reliability index of around 0.6 but not less than 0.5 was accepted and considered to be useful to differentiate students on the basis of estimated abilities, where values less than 0.5 were ascribed to measurement error (Fisher, 1992; Griffin, Care, and Harding, 2015). Nonetheless, observed item separation reliability of 0.992 and person separation reliability of 0.6 were used as an evidence of construct as well as criterion validity (Wright and Masters, 1982). This provided validity that these tasks could be used for estimating ability parameters of those students who completed the task.

Interpretation and Developing a Skills Progression

After calibration, the estimates of the indicator difficulty parameters (δ) were arranged in ascending order and interpreted in order to see whether they defined the hypothesised conceptual framework. The conceptual framework (Callingham, 2004; Griffin, 2014) was examined by using a variable map, which showed the grouping and clustering of different indicators along the unidimensional scale. Clusters of indicators were further explored to see whether these clusters could together be interpreted as skill level. Wright and Masters (1982) proposed that each item could help define the variable underpinning the construct by separating the items based on the description of underlying skills. Each indicator was interpreted as students' behaviours. This clustering was according to the similar difficulty level of indicators where the ability of students matched with the difficulty of the item. Hence, both are mapped onto the same scale. In order to establish a substantive interpretation of the scale, a similar analysis was undertaken, as described by Callingham (2004), Griffin (2007a, 2014),

Griffin et al. (2014), and Woods (2010).

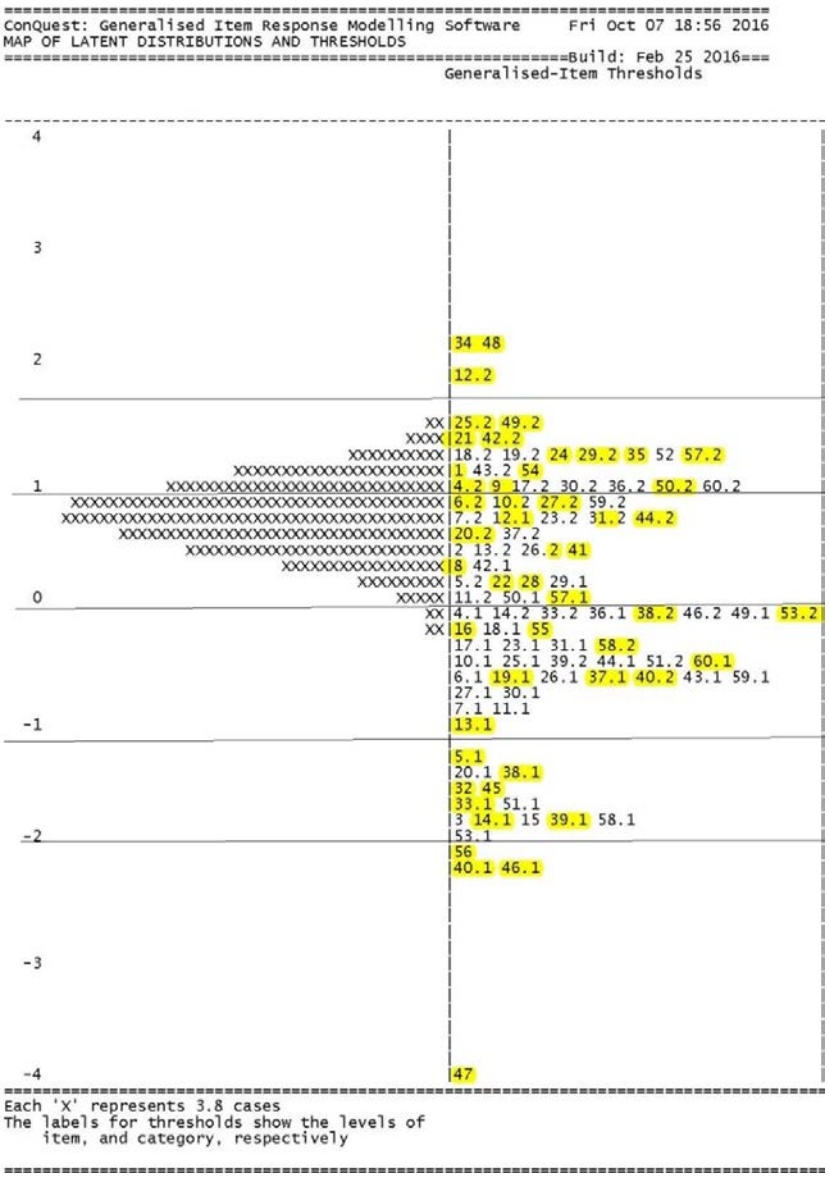

Figure 8. Variable map for a set of nine tasks as assessment instrument and selected indicators along the continuum.

Tables 6 (A–E) includes all indicators and their interpretations as indicative behaviours. The cut points are based on the change of difficulty as discussed above. A summarised interpretation was produced to describe in brief the quality criteria needed to respond successfully to items appearing in the same zone of the problem-solving scale. The interpretation appears to be reliable with an underlying theorised continuum of competence.

Table 6A. Level A (1) Item Criteria, Delta Parameters and Derived Standard for interactive problem-solving level One

Criterion		Thresholds	Rubrics	Interpretation	Derived Standard
56.1	OO2L_VF	-2.109	Goal directed	Although it is seen that they are starting with goal-directed actions but there is no more evidence	**Nonparticipation/guessing:** At this level, there is insufficient information to
40.1	DSt_RoC	-2.156	Confused/ Distracted/Uncertainty		

46.1	DDT_RoC	-2.156	Confused/ Distracted/Uncertainty	except that some of them are submitting the correct solution, it is because of random interaction with the task.	describe the student's problem-solving ability. If the student interacted with the problem, the structure of the task was apparently not understood.
47.1	REV_Corr	-3.891	Guessing		

Table 6B. Level B (2) Item Criteria, Delta Parameters and Derived Standard for interactive problem-solving level Two

Criterion		Thresholds	Rubrics	Interpretation	Derived Standard
5.1	OO4L_RT	-1.305	Time is intermediately enough to explore.	The time taken to explore is very limited and not enough to explore the whole problem space, but some evidence of systematic exploration and random trial-and-error attitude, along with some errors, shows that they are exploring the problem space. However, their uncertain attitude of rolling the mouse for submission indicates that they may exit or leave the task at any time without solving the problem.	**Exploring:** Students at this level start exploring the individual elements within the task in a random manner. Unsystematic trial-and-error approaches are used in an attempt to solve the problem. They may exit the task without solving the problem.
38.1	DSt_AC	-1.328	Random trial and error		
20.1	BTR_RoC	-1.414	Minimal uncertainty/ Some distraction		
32.1	DDO_SS	-1.508	Systematic exploration		
51.1	REV_INC	-1.625	Some errors		
3.1	OO4L_VF	-1.789	Goal directed		
39.1	DSt_SS	-1.813	Random exploration		

Table 6C. Level C (3) Item Criteria, Delta Parameters and Derived Standard for interactive problem-solving level Three

Criterion		Thresholds	Rubrics	Interpretation	Derived Standard
				The latencies and extended time for exploration as well as completion shows that students are trying to recognise elements. A mix of random and systematic trials leads to the evidence that they are trying to find the patterns and connections between these elements. Their complete focus on the task and task instruction along with trailing of some solution strategies illustrates that they have found the patterns to follow. Evidence of some planning found, but not executed or unable to execute. The systematic trials show an elimination of invalid	**Recognising** patterns: Students at this level identify the elements within the task and explore possible solutions. They recognise and use patterns of events to help direct them towards the solution. They search for connections between different elements of the problem.
53.2	REV_RoC	-0.023	Completely Focused/confident with no uncertainty		
49.1	REV_RT	-0.039	Time is intermediately enough to explore.		
38.2	DSt_AC	-0.047	Systematic trial and error		
4.1	OO4L_PSL	-0.063	Limited efforts and time to understand the problem		
33.2	DDO_RoC	-0.063	Completely Focused/confident with no uncertainty		
36.1	DSt_PSL	-0.063	Limited efforts and time to understand the		

			problem	actions along with a comprehensive search of the problem space.	
55.1	OO2L_BF	-0.141	Following instruction		
16.1	BTR_BF	-0.148	Following instruction		
18.1	BTR_PSL	-0.211	Limited efforts and time to understand the problem		
31.1	DDO_AC	-0.313	Moderately efficient/random trial and error		
44.1	DDT_AC	-0.43	Moderately efficient/random trial and error		
51.2	REV_INC	-0.434	Flawless interactions/Error avoidance		
39.2	DSt_SS	-0.504	Checking some of the possible solution		
6.1	OO4L_AC	-0.641	Moderately efficient/random trial and error		
11.1	'BSt_RT'	-0.852	Time is intermediately enough to explore.		
50.1	REV_PSL	0.016	Limited efforts and time to understand the problem		

Table 6D. Level D (4) Item Criteria, Delta Parameters and Derived Standard for interactive problem-solving level Four

Criterion		Thresholds	Rubrics	Interpretation	Derived Standard
17.2	BTR_RT	1.047	Optimum time taken to plan and execute the goal appropriately.	The students at this level seem familiar with all the elements of the task as evident from their systematicity and efficiency. They are trying to connect these elements by starting with the most goal-directed actions according to the instruction, and their focus on planning and execution is evident from their planning time that they have put optimum effort to deep understanding of the task. Completion time is optimum to execute their plan. There is evidence of some errors along with	**Forming rules:** Students at this level ensure they are familiar with all the elements of the problem. They search for patterns to connect elements of the problem, recognising repeating elements, steps, or stages and building these into rules for solving the problem.
50.2	REV_PSL	1.047	Optimum efforts and time/deep understanding of the problem		
9.1	BSt_VF	0.984	Most goal directed		
36.2	DSt_PSL	0.945	Optimum efforts and time/deep understanding of the problem		
27.2	LWC_RoC	0.902	Completely Focused/confident with no uncertainty about their		

			solution strategy	some randomness, which might be evidence of their trials to recognise some repeated elements in the tasks. Systematicity and efficiency along with some random trials suggests it is most likely that they are forming some rules to follow.	
6.2	OO4L_AC	0.867	Systematic trial and error		
7.2	OO4L_RoC	0.734	Completely Focused/confident with no uncertainty about their solution strategy		
12.1	BSt_INC	0.695	Some errors		
13.2	BSt_AC	0.531	Systematic trial and error		
26.2	LWC_AC	0.531	Systematic trial and error		
2.1	OO4L_BF	0.414	Planned/clear understanding of the task/following instruction		
8.1	BSt_Corr	0.375	Goal achieved		
5.2	OO4L_RT	0.219	Optimum time taken to plan and execute the goal appropriately		
57.1	OO2L_AC	0.102	Random trial and error		

Table 6E. Level E (5) Item Criteria, Delta Parameters and Derived Standard for interactive problem-solving level Five

Criterion		Thresholds	Rubrics	Interpretation	Derived Standard
49.2	REV_RT	1.523	Optimum time taken to plan and execute the goal appropriately	The highly planned, systematic, and efficient behaviour along with all correct solution leads to the evidence that they are applying/testing the rules identified in the previous level. There is no evidence of errors or randomness at this level. It is depicted by the searches and actions that are necessary to solve the problem.	**Testing rules and forming hypotheses:** Students at this level form rules and test them systematically. They think ahead to avoid errors and ensure actions contribute to the solution.
21.1	LWC_Corr	1.43	Goal achieved		
24.2	LWC_SS	1.414	Systematically trials and testing of all possible solutions		
19.2	BTR_AC	1.313	Highly efficient/systematic trial and error		
52.1	REV_AC	1.297	Highly efficient/systematic trial and error		
57.2	'OO2L_AC'	1.297	Highly efficient/systematic trial and error		
35.1	DST_BF	1.273	Planned/clear understanding of the task/following instruction		

29.2	DDO_PSL	1.234	Optimum efforts and time/deep understanding of the problem		
54.1	OO2L_Corr	1.188	Goal achieved/correctness		
43.2	DDT_RT	1.137	Optimum time taken to plan and execute the goal appropriately		
1.1	OO4L_Corr	1.125	Goal achieved/correctness		

Table 6F. Level F (6) Item Criteria, Delta Parameters and Derived Standard for interactive problem-solving level Six

Criterion		Thresholds	Rubrics	Interpretation	Derived Standard
48.1	REV_BF	2.25	Planned/clear understanding of the task/following instruction	Shows high level of problem-solving skills that demand higher-order thinking as they are predicting based on their previous outcomes. They identified in previous levels that whenever they spent time to plan their solution properly, their very first goal-directed and solution-oriented step (based on the task instruction) would lead them to the correct solution. Therefore, at this level their solution strategy appeared to be flawless and efficient. They are indeed trying to apply and generalise those rules they developed at previous levels. At this stage, the students are expected to learn how to judge the quality of their learning process. Detailed planning is evident, with ability to execute what was planned.	**Making generalisations for transfer:** Students at this level solve tasks systematically. They are able to make generalisations so that the strategies learned in one task can be applied to other problems.
34.1	DSt_Corr	2.211	Gaol achieved/ correction		
12.2	BSt_INC	1.969	Flawless solution		

Six levels of proficiency were thus identified and summarised for the measure of problem-solving ability. Items clustered at the lowest level did not have sufficient information to describe the student's problem-solving ability. At the first level, if the student interacted with the problems, the structure of the task was apparently not understood. In addition, students at this stage closed the problem prematurely, as evident by their failing to gather enough information to solve the problem decisively. Therefore, their efforts seemed to be at the level of guesswork. At the second level, items described the capacity of students to explore the individual elements within the task in a random manner. They were now able to explore the problem space by using unsystematic trial-and-error approaches. The

third level represented the students' development of the ability to identify the elements within the task and explore possible solutions. They recognised and used patterns of events to help direct them towards the solution. Items that clustered at the fourth level denoted the development of proficiency to search for patterns to connect elements of the problem, recognise repeating elements, steps, or stages and to build these into rules. Those at the fifth level represented students' capacity to form rules and test them systematically, with items at the sixth and highest level representing students' development of the ability to solve tasks systematically and make generalisations so that the strategies learned in one task could be applied to other problems.

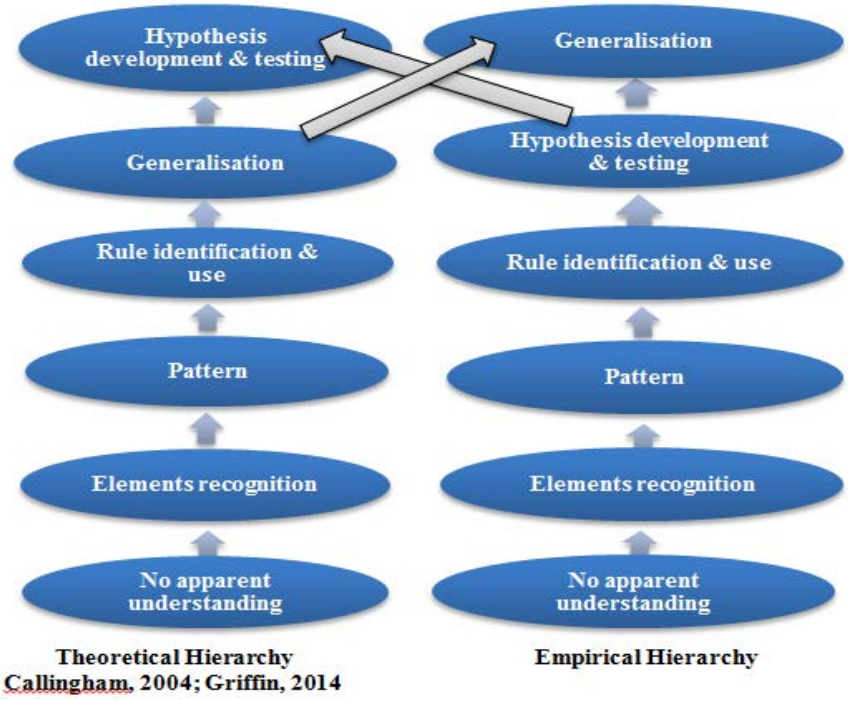

Figure 9. Comparison between theoretical and empirical continuum.

There is a very close match found between the conceptual framework and the empirically derived framework with the exception of the highest two levels, which appeared to be a switch between generalisation and hypothesis development and testing ability (Figure 9). Thus, the hypotheses of this study were confirmed. The psychometric model (IRT) was successfully fitted to the available data (Griffin et al., 2006) obtained from problem-solving game-based tasks. In other words, this conceptual model (Callingham & Griffin, 2000, 2001; Griffin, 2014) best fits the indicators extracted from the log stream file within the psychometric model.

Summary and Conclusion

This paper endeavoured to explore a methodology to infer students' behavioural and cognitive processes from log stream data that were collected during their performance on complex tasks (i.e., computer games). The primary aim of the study was the design and validation of measures obtained from game-based tasks of problem-solving, and the inferences drawn from students' performance. The study attempted to improve on some limitations of problem-solving assessment by linking the computer interactions of students with developmental assessment. A set of indicators were identified and redefined from the secondary data, and scoring rules were developed through which this linking was accomplished.The Rasch partial credit model (Masters, 1982) was used to obtain a measurement scale, and the item clusters produced by the Rasch analysis were subjected to a skills audit and then translated onto the underlying continuum. The fit to the model was acceptable across all tasks and items (.96 to 1.06), and given the exploratory nature of data these fit statistics were considered to be

within an acceptable range and provided evidence that the behavioural indicators or items that were used to measure problem-solving could define a single dimensional latent construct. The data obtained from this process satisfied most of Messick's (1989) criteria for construct validity. The most innovative part of the study was that the linking of computer interactions with developmental assessment provided ways of developing performance assessments that could be generalised across different subjects, grade levels, and over time.

It is recommended that this construct could be drawn by the teachers in the classroom, generating their own data instead of interactive data. There are underdeveloped countries that are struggling to equip their learning with the modern technology and may feel left behind. However, teachers can draw and use this construct by investigating different forms of indicators as evidence by designing different problem-solving games for assessment and teaching. In the classroom situation, it is possible for teachers to design tasks that are interesting that are also based on real-life problems in order to engage students fully and positively, and to use this construct to assist students in improving their problem-solving skills. This study, therefore, recommends that when a learning progression of interactive problem-solving is used in combination with targeted teaching, considerable improvements in student problem-solving skills and in teacher knowledge and behaviour could be attained.

This paper describes the development of a general progression of cognitive development in the field of problem-solving. Given that there appeared to be limited consensus in the literature as to the definition and dimensionality of generic (rather than domain specific) problem-solving , an inductive approach was appropriate. In addition, the generic conceptual framework and Rasch modelling both assumed unidimensionality, which limited the possibility for multidimensional models of cognitive development/ competence to be tested. However, "if the items can be shown to be systematically and predictably related to each other along the variable, this confirms that a single construct is being measured, and provides evidence of construct validity" (Callingham, 2004, p.122). Thus, in the context of this study, fit indices were examined to investigate the degree of correspondence between the conceptual model and the measurement model predicted by the Rasch model. In addition, a further study was conducted to establish criterion validity and to monitor the predictive validity; however, some quasi experimental studies could be needed in future.

Two important activities which can be recognised as future directions for this research are the improvement of the problem-solving construct and intervention strategies in the classroom. Firstly, improving the problem-solving construct entails additional validation studies, continued task constructions and tasks with additional problem space in order to trace further specific indicators. Further validation is required due to the narrow range of sample ability variance (.102) and the low person seperation reliaibility (.6) possibly because of the relatively few indicators, most of which were common across the tasks measuring the same behaviour. Additionally, the research was limited to secondary data which have major limitations inherited in its nature since this type of data was only meant to answer specific research questions restricted to certain contexts. There were some elements that were beyond the control of the researcher, including task design, number of task and indicators, sample design and size. In addition, the inferences drawn from students' interactions were based on the researcher's personal judgements and understanding based on the theoretical background. It was impossible for the researcher to check the consistency of the rubrics execution and interpretation across raters due to cost involved. Secondly, intervention strategies such as implementation of the construct in the classroom to determine the effectiveness of the developmental progression in improving students' problem-solving ability is recommended.

In summary, use of this developmental progression, if shown significant improvements in the development of higher-order thinking skills, could be used at different levels of schooling for the enhancement of students' problem-solving ability.

References

Adey, P., Csapó, B., Demetriou, A., Hautamäki, J., & Shayer, M. (2007). Can we be intelligent about intelligence? Why education needs the concept of plastic general ability. *Educational Research Review 2*, 75–97.

Assessment Research Centre. (2004–2014). *Assessment and learning partnerships (ALP) project*. Retrieved from http://education.unimelb.edu.au/arc/projects/completed/2014/alp

Assessment Research Centre. (2007–2010). *Students with additional needs (SWAN) project*. RetrievedJuly 15, 2016 from http://education.unimelb.edu.au/arc/projects/completed/swans

Assessment Research Centre. (2009–2012). *Assessment and teaching of 21st century skills (ATC21s) project*. Retrieved from http://education.unimelb.edu.au/arc/projects/completed/atc21s

Assessment Research Centre. (2009-2013) *ABLES project*. Retrieved from http://education.unimelb.edu.au/arc/projects/abilities-based-learning-and-education-support-ables-research

Assessment Research Centre. (2011). *Assessment and learning partnerships*. Retrieved from http://education.unimelb.edu.au/arc/projects/completed/2014/alp

AusVELS. (2014). *The Australian Curriculum in Victoria.* Retrieved from http://ausvels.vcaa.vic.edu.au/

Adams, R., Vista, A., Scoular, C., Awwal, N., Griffin, P., & Care, E. (2015). Automatic coding procedures. In P. Griffin and E. Care. (Eds.) *Assessment and teaching of 21st century skills: Methods and approaches* (pp. 115–132). Dordrecht, The Netherlands: Springer.

Biggs, J. B. & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York, NY: Academic Press.

Biggs, J. B. & Collis, K. F. (1991). Multi-modal learning and the quality of intelligent behaviour. In: H. Rowe (Ed.). *Intelligence: Reconceptualisation and measurement.* Hillsdale, NJ: Lawrence Erlbaum.

Bloom, B. (1956). *Taxonomy of educational objectives: The classification of educational goals*. London, England: Longman.

Boud, D. and Falchikov, N. (2007). *Rethinking assessment in higher education. Learning for the longer term*. Hoboken, NY: Taylor and Francis.

Callingham, R. A. (2004). *Numeracy assessment: From functional to critical practice* (Published doctoral dissertation). Faculty of Education, The University of Melbourne, Melbourne, Australia.

Callingham, R. & Griffin, P. (2000). Towards a framework for numeracy assessment. In J. Bana & A. Chapman (Eds.). *Mathematics Education beyond 2000. Proceedings of the 23rd Annual Conference of the Mathematics Education Research Group of Australasia* (pp. 134–141). Fremantle, Australia: MERGA.

Callingham, R. & Griffin, P. (2001). Shaping assessment for effective intervention. In *Mathematics Shaping Australia*. *Proceedings of the 18th Biennial Conference of the Australian Association of Mathematics Teachers* [CDROM]. Canberra, Australia: AAMT.

Chi, M. T. H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by expert and novices. *Cognitive Science, 5*. 121–152.

Csapó, B. (2007). Research into learning to learn through the assessment of quality and organization of learning outcomes. *The Curriculum Journal. 18*(2), 195–210.

Crisp, G. (2010). Interactive E-Assessment--Practical Approaches to Constructing More Sophisticated Online Tasks. *Journal of Learning Design*, *3*(3), 1-10.

Dreyfus, S., & Dreyfus, H. (1980). *A five-stage model of the mental activities involved in directed skill acquisition*. Washington, DC: Storming Media.

Fisher, W. Jr. (1992). Reliability statistics. *Rasch Measurement Transactions, 6*(3), 238.

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

Funke, J. (2010). Complex problem solving : A case for complex cognition? *Cognitive Processing, 11*, 133–142.

Gillis, S., & Griffin, P. (2004). Using rubrics to recognise varying levels of performance. *Training Agenda: A Journal of Vocational Education and Training, 12*(2), 22–24.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 118*, 519–521.

Griffin, P. (2000). *Competency based assessment of higher order competencies*. Paper presented at the NSW ACEA State Conference, Mudgee, Australia.

Griffin, P. (2001). *Performance assessment of higher order thinking*. Paper presented at the annual conference of the American Education Research Association, Seattle.

Griffin, P. (2007a). The comfort of competence and the uncertainty of assessment. *Studies in Educational Evaluation, 33*, 87-99.

Griffin, P. (2007b). *Evidenced based teaching and curriculum shifts.* Assessment Research Centre,The University of Melbourne. Retrieved from https://www.google.com.au/search?q=Griffin,+McGaw,+%26+Care,+2012&ie=utf-8&oe=utf-8&gws_rd=cr&ei=WPDEV7PiJITM0ATW-6_4DA#q=Evidenced+based+teaching+and+curriculum+shifts

Griffin, P. (2009). Teacher's use of assessment data. In C. Wyatt-Smith & J. Cumming (Eds.). *Educational assessment in the 21st century*. Dordrecht, The Netherlands: Springer.

Griffin, P. (2014). Performance Assessment of Higher Order Thinking. *Journal of Applied Measurement, 15 (1)*: 53–68

Griffin, P., Care, E., Francis, M., Hutchinson, D., Awwal, N., Mountain, R., Pavlovic, M., Robertson, P., & Woods, K. (2014). *Assessment for Teaching*. Melbourne, Australia: Cambridge University Press.

Griffin, P., Care, E. & Harding, S. (2015). Task characteristics and calibration. In P. Griffin and E. Care (Eds.). *Assessment and teaching of 21st century skills: Methods and approach*. (pp. 133-182). Dordrecht, The Netherlands: Springer Science and Business Media.

Griffin, P., Care, E., Robertson, P., Crigan, J., Awwal, N., & Pavlovic, M. (2013). Using developmental framework of learning. *Assessment and learning partnerships in an online environment*: GI Global.

Griffin, P., Mak, A., & Wu, M. L. (2006). *LP0561852: Modelling person and item parameters of computer administered problem-solving strategies:* ARC application. Melbourne, Australia: Melbourne University Research Office.

Griffin, P. E., McGaw, B., & Care, E. (2012). *Assessment and teaching of 21st century skills*, In P. Griffin, B. McGaw, E. Care (Eds.). Dordrecht, The Netherlands; New York, NY: Springer.

Griffin, P. & Robertson, P. (2014). Judgement-based assessment. In P. Griffin (Ed.), *Assessment for teaching*, (pp. 107–124). Melbourne, Australia: Cambridge University Press.

Greiff, S., Holt, D. V., & Funke, J. (2013). Perspectives on problem solving in educational assessment: Analytical, Interactive, and Collaborative Problem Solving . *The Journal of Problem Solving , 5(2)*, 71–91.

Hesse, F., Care, E., Buder, J., Sassenberg, K., & Griffin, P. (2015). A framework for teachable collaborative problem solving skills. In P. Griffin and E. Care (Eds.). *Assessment and teaching of 21st century skills: Methods and approaches. Educational assessment in an information age.* (pp. 37–56). Dordrecht, The Netherlands. Springer Science and Business Media.

Hutchinson, D, Francis, M., & Griffin, P. (2014). Developmental teaching and assessment. In P. Griffin (Ed.), *Assessment for teaching*, (pp. 26–57) Melbourne, Australia: Cambridge University Press.

Jonassen, D. (2000). Toward a design theory of problem solving . *Educational Technology Research and Development*, *48* (4), 63-85.

Krathwohl, A., Bloom, B., & Masia, B. (1964). *Taxonomy of educational objectives; The classification of educational goals.* Handbook II: The affective domain. New York, NY: Longman Green.

Limjap, A. A. (2011). An analysis of the mathematical thinking of selected Filipino pupils. *Asia-Pacific Education Researcher*, *20*(3), 521-533.

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149- 174.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Organisation for Economic Co-operation and Development (OECD). (2003). *The PISA 2003 assessment framework. Mathematics, reading, science and problem solving knowledge and skills*. Paris, France: OECD.

Organisation for Economic Co-operation and Development (OECD). (2007). *PISA 2006: Science competencies for tomorrow's world. Volume 1: Analysis.* Retrieved from http://www.pisa.oecd.org/dataoecd/30/17/39703267.pdf

Organisation for Economic Co-operation and Development (OECD). (2010). *The PISA 2012 field trail problem solving assessment framework.* Draft subject to possible revision after the field trial. Paris, France: OECD.

Organisation for Economic Co-operation and Development (OECD). (2013). *PISA 2012 Assessment and analytical framework*. Retrieved from http://www.oecd.org/pisa/pisaproducts/PISA%202012%20framework%20e-book_final.pdf

Organisation for Economic Co-operation and Development OECD. (2013). *PISA 2015 draft collaborative problem solving framework*. Paris, France: OECD.

O'Neil, H. F. Jr (1999). Perspectives on computer-based performance assessment of problem solving , *Computers in Human Behavior,* 15 (3/4), 225–268.

O'Neil, H. F. Jr. (2002). Perspectives on computer-based performance assessment of problem solving . *Computers in Human Behavior*, 18(6), 605–607.

O'Neil, H. F. Jr., Sabrina, S.H.C., & Gregorky. W. K. C. (2003). Issues in the computer-based assessment of collaborative problem solving . *Assessment in Education, 10(3),* 361–373.

Ohlsson, S. (2012). The problems with problem-solving : Reflections on the rise, current status, and possible future of a cognitive research paradigm. *The Journal of Problem Solving, 5(1),* 101–128.

Partnership for 21st Century Skills. (2009, May). *P21 framework definitions document*. Retrieved from http://www.p21.org/our-work/p21-framework

Pólya, G. (1973). *How to solve it: A new aspect of mathematical method.* Princeton, NJ: Princeton University Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danish Institute for Educational Research.

Schoenfeld, A. H. (1985). *Mathematical problem solving* . San Diego, CA: Academic Press.

Sternberg, R. J. (1996). What is mathematical thinking? In R. J. Sternberg & T. Ben-Zeev (Eds.), *The nature of mathematical thinking* (pp. 303–318). Mahwah, NJ: Lawrence Erlbaum Associates.

Tricot, A. & Sweller, J. (2014). Domain-specific knowledge and why teaching generic skills does not work. *Educational Psychology Review*, *26*(2), 265–283.

Vygotsky, L. (1978). Mind and society: The development of higher psychological processes. Cambridge, MA: Harvard University Press.

Wu, M. L. (2003). *The application of item response theory to measure problem-solving proficiencies.* (Published doctoral dissertation). The University of Melbourne, Melbourne, Australia.

Wu, M., & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*. Melbourne: Educational Measurement Solutions.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ConQuest: Generalised item response modelling software*. Camberwell, Australia: Australian Council for Educational Research.

Wolfe, E. W., & Smith, E. V. Jr. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II—Validation activities. In E. V. Smith, Jr. & R. M. Smith (Eds.). *Rasch measurement: Advanced and specialized applications* (pp. 243–290). Maple Grove, MN: JAM Press.

Woods, K. (2010). *The design and validation of measures of communication and literacy to support the instruction of students with learning disabilities.* (Published doctoral dissertation). The University of Melbourne, Melbourne, Australia.

Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis.* Chicago, IL: MESA Press.

Zoanetti, N. (2009). *Interactive computer based assessment tasks: How problem-solving process data can inform instruction*. (Published doctoral dissertation). The University of Melbourne, Melbourne.