ACTA IMEKO
ISSN: 2221-870X
September 2024, Volume 13, Number 3, 1 - 7

A validity exploration model related to the existence of the generic problem-solving competence

L. A. Nguyen Khoa[1], M. Courtney[2], M. Wilson[3], C. T. K. Nguyen[1]

[1] *University of Melbourne, 100 Leicester Street, Melbourne, Australia*
[2] *Nazarbayev University, Kazakhstan*
[3] *University of California, Berkeley, USA*

ABSTRACT
There has been a growing focus on exploring the existence of Generic Problem-Solving competence across various fields, leading to heightened attention in this area. However, most of the previous and current approaches are limited in terms of validity and reliability. Thus, this paper aims to propose a new approach based on The Standards for Educational and Psychological Testing validity framework to investigate this matter. The investigation leads to the review of the conceptions of construct validity in educational measurement. The paper concentrates on the proposed validity exploration (VE) model, representing an elaborate enterprise and a serial, progressive procedure aligned with the content and structural validity aspects of The Standards framework. The PISA Computer-based Assessment was used as secondary data for this investigation.

Section: RESEARCH PAPER

Keywords: Generic; Problem-Solving; PISA; validity; reliability

Citation: L. A. Nguyen Khoa, M. Courtney, M. Wilson, C. T. K. Nguyen, A validity exploration model related to the existence of the generic problem-solving competence , Acta IMEKO, vol. 13 (2024) no. 3, pp. 1-7. DOI: 10.21014/actaimeko.v13i3.1338

Section Editor: Eric Benoit, Université Savoie Mont Blanc, France

Received July 11, 2022; **In final form** September 4, 2024; **Published** September 2024

Funding: This work was supported by The University of Melbourne, Australia.

Corresponding author: L. A. Nguyen Khoa, e-mail: la.nguyenkhoa@unimelb.edu.au

1. BACKGROUND AND OBJECTIVE OF THE STUDY

Common attention for investigators, politicians, teaching staff, and the community is if performances in measures of domain-specific problem-solving (PS) and domain-general PS should address a common generic core of PS competence [1]-[4]. Instances of correlations between students' performance in both domain-specific PS and domain-general PS and the evidence of several generic competencies across disciplines [5], as well as the similarities across PS abilities in different contexts [6], [7], make it important to identify the presence (or not) of a latent Generic PS scale underlying both domain-specific and domain-general PS scales. Thus, more investigation should be conducted on the existence of a Generic PS construct, and how it cuts across disciplines. The need is more urgent because the measured domain-specific PS constructs and domain-general PS constructs are required to be associated since they all include intellectual capabilities essential for producing as well as implementing regulations and they are also foreseen to show singular variance in a substantial way.

Nevertheless, comprehensive studies are scarce on the existence of a Generic PS construct underlying both domain-specific PS constructs and domain-general PS construct [8]. Intending to explore the presence of the latent Generic PS scale across both measures, most studies have compared certain different PS constructs, which provided evidence supporting some common generic core PS skills underlying constructs. However, they were somewhat limited in terms of (i) the extent of the content analysis undertaken, or (ii) the psychometric methodologies applied to investigate the existence of Generic PS across both domain-specific and domain-general PS constructs.

The current study, proposing a validity exploration (VE) model to explore the existence of the Generic PS construct (using quantitative methods), is designed to tackle some of the aforementioned shortcomings. The VE model was based on The Standards framework and the validity argument model by Chapelle, Enright, and Jamieson [9]. In particular the VE model examines the existence of the Generic PS construct with respect to two domain-specific PS constructs and one domain-general PS construct measured by the PISA 2012 CBA tests. This study

includes perspectives from mathematics, reading (the two domain-specific subjects) and creative PS (the single domain-general subject), drawing upon a culturally diverse and educationally homogenous sample of 15-year-old students. This complies with the advice of Alderson [10], who strongly suggested that researchers' study learners whose education and culture is reasonably homogenous, so as to better understand the character of PS competence.

2. LESSONS FROM PREVIOUS APPROACHES FOR ASSESSING THE EXISTENCE OF A GENERIC PROBLEM-SOLVING CONSTRUCT ACROSS DOMAIN-SPECIFIC AND DOMAIN-GENERAL PROBLEM-SOLVING CONSTRUCTS

Table 1 below summarises studies that were identified as exploring the existence of the Generic PS construct underlying both domain-specific and domain-general PS constructs. These studies are classified methodologically and discussed in the context of their validity and reliability.

With regard to the use of qualitative methods, the author could identify only one investigation which focussed on the existence of Generic PS constructs underlying both domain-specific and domain-general PS constructs, the Baird [5] study. However, Baird did not make use of subject matter experts to come up with these classifications of items. In this instance, only the author acted as an expert. Therefore, Baird's [5] study is somewhat lacking in evidentiary strength as the findings were not corroborated. Messick [11] pointed out that a judgement which is made by one researcher may be subject to unreliability, bias, and error. In addition, Baird's [5] early study did not account for assessments that employed computer-based simulation problems which allow test administrators to administer tests in an efficient manner. Baird concluded that although there are PS schemes and capabilities which are in use frequently across disciplines, they still depend on an understanding of the field and the processes required to perform them, and they are mostly established by each field [5]. Baird also pointed out that it is uncertain if these capabilities expand from one kind of subject matter to another kind within a domain and if they can be actually taught to beginners in the domain.

With regard to the quantitative methods, the most common methodology in research on the existence of a Generic PS construct underlying domain-specific as well as domain-general PS constructs is the use of confirmatory factor analyses [3], [6], [7], [12]. To the best of the researchers' knowledge, item response theory (IRT) has hardly ever been used in investigations in this field. Although Rasch models were used in the Molnár et al. [8] and Molnar et al. [13] studies, they are used only for scaling the data, not for modelling the relationship between domain-specific and domain-general PS constructs. Briggs and Wilson [14] argue that the reason for this used to be the statistical problems associated with fitting IRT models and the complications related with interpreting the resultant variables. There are some potential drawbacks of the Confirmatory Factor Analysis (CFA) approach used in the previous studies in this field which result in poor reliability. One of the limitations is that the item traits and student characteristics cannot be separated. As a result, the difficulty of items depends on the students' ability, which violates the assumptions of objective measurement. However, reliability is the central concept in this true score theory [15]; it seems that CFA's reliability assumption might not be precise in true situations. Under some situations, the students are not measured with the same degree of accuracy because of differences in their ability levels and the test difficulty. As a result, these results are not replicable [16]. Thus, previous studies have concluded that additional study needs to consider using models of IRT as well as testlets as methodological approaches for the purpose of more precisely analysing the construction of PS competence [7], [17]. In addition, as De Boeck and Wilson [18] emphasised that IRT models receive more and more attention in psychology and epidemiology.

Furthermore, most of the studies on the existence of Generic PS underlying both domain-specific and domain-general PS constructs have been conducted with one national population, such as the Hungarian studies by [8] and [13] and the German studies by Scherer and Tiemann [6], [17]. In addition, none of the studies compared the PS constructs across three distinct PS constructs. In all the related research projects listed above, the focus was on comparing PS capability between one domain-specific and one domain-general PS construct [3], [6], [[7], [8], [13], [17]. Thus, to date, no existing research has looked at the relationship of PS constructs across more than two areas.

3. THEORETICAL FRAMEWORK

As previously mentioned in Section 2, most studies assessing the existence of the generic PS construct underlying both domain-specific and domain-general PS constructs have shortcomings related to validity. Thus, this section aims to examine relevant published writings concerned with issues of validity and validation. By this way, this section provides the foundation for the proposed validation exploration model in this study. Topics to be covered include both the development of the concepts of validity and a framework for collecting validation evidence in educational measurement: Early definitions of validity [12], [19], [20], Messick [21], [22]'s framework for the unified concept of construct validity, and the Standard for Educational and Psychological Testing [1], [2].

3.1 Early definitions of validity

Validity has been divided into many different types, such as "face validity, validity by definition, intrinsic validity, logical validity, empirical validity, factorial validity, etc". [23], p2. However, the definitions of validity have not always been generally accepted. Traditionally, the conception of validity began as a criterion-based model related to a content model before the construct model emerged [19]. In the early years, validity used to be defined simply as the appropriateness and accuracy of test score interpretations [12]. Kelley's [24] conception of validity was representative of this stage (as cited in 25, p. 1061). However, his definition is simplistic as he simply argued that "a test is valid if it measures what it purports to measure" [24], p.14. As a result, a few researchers called for another conception of validity that is based more on the criterion model [25]. According to Angoff [26], this model is credited to Cureton [20] who described the criterion model of validity. He provides an operational definition of validity as the "correlation of observed scores on the test with true scores on the criterion" [26] p. 20 and discusses the differences in a test's validity, predictive power, and relevance. It is clear that Cureton placed emphasis on the criterion and reflected the general thinking of the time (see [26]).

The construct dimension of the definition of validity was presented by Cronbach and Meehl [27]. They suggested that construct validity would be required once an examination is explained as an assessment of some competencies, which are not determined or do not have an adequate criterion. Thus, in the

Table 1. Summary of Different Approaches for Examining the Existence of Generic PS Construct Underlying Domain-Specific and Domain-General PS Constructs

Approaches	Source of data	Participants	Statistical Technique used	Findings	Studies
Examining domain-specific PS skills that are common across fields	Assessing documents from previous studies on domain-specific PS	The author	Qualitative method: comparison of domain-specific PS skill descriptions seeking similarities between skills	Similar skills are used in different fields.	Baird (1983)
Examining the common parts of most of the PS processes	Assessing documents from previous 11 studies from 1972 to 1998	Two researchers	Literature review: summarising the common parts about cognitive and metacognitive aspects of PS	There are six component parts of most of the PS processes	Wirth and Klieme (2003)
Examining the relationship between dynamic PS and analytic PS	Collected item-response data.	654 German students from 18 different types of schools	Quantitative method: Correlation coefficients estimated by way of Structural equation models	It emerged that analytical and dynamic aspects of PS have to be distinguished	Wirth and Klieme (2003)
Examining the relationship between analytic and complex PS in the structure of PS competence		162 students from Grade 10 and the upper secondary level	Confirmatory factor analyses: used to establish measurement models representing different theoretical assumptions	Both analytic and complex PS are distinct constructs.	Scherer and Tiemann (2012)
Examining the relationship between domain-general scientific PS (complex-interactive PS) construct and domain-specific scientific PS construct (analytical-static PS)	Collected item-response data.	805 German high-school students from Grades 8 and 10	Confirmatory factor analysis: to establish a measurement model	There are common cognitive processes shared in different contents and contexts. The processes involved in domain-specific PS are comparable to those in domain-general PS	Scherer and Tiemann, (2014)
Examining the relationship between analytic PS and interactive PS	Collected item-response data	339 university students and 577 high-school students	Quantitative method: Latent regression/correlations and an analysis of commonalities	Both PS address a common core of PS competency. Both PS construct are highly interrelated	Fischer et al. (2015)
Examining the relationship between complex PS (domain-general) and domain-specific PS (three types of problems based on the amount of information given).	Collected item-response data	600-800 students from 3rd to 11th grade (aged 9-17) in Hungarian primary and secondary schools.	Quantitative method: Using internal consistencies of the tests (Cronbach's alpha) and bivariate correlations	The correlation between domain-specific and complex PS proved to have increased over time. The constructs are related but do not constitute the same construct.	Molnár, Greiff, and Csapó (2013)
Examining the relationship between domain-specific PS (static scenarios, mathematics and science) and domain-general PS (interactive scenarios, complex PS)	Collected item-response data	788 students from 5th to 11th grade (aged 11-17) in Hungarian primary and secondary schools	Structural Equation Model: The bivariate correlations	The correlation between domain-specific and complex PS proved to have increased over time. The constructs are related but do not constitute the same construct.	Molnár, Greiff, Wüstenberg, and Fischer (2017)

the article "Construct validity in psychological tests", they introduced four types of validation. They were predictive validity, concurrent validity, content validity, and construct validity [27]. Cronbach and Meehl [27] claim that predictive and concurrent validity should take into account validity procedures based on a criterion or be criterion oriented.

3.2 Messick's Framework for the unified concept of construct validity

In 1957, Loevinger argued that "since predictive, concurrent, and content validities are all essentially adhoc, construct validity is the whole of validity from a "scientific point of view" [28] p. 636. From here, the construct model of validity has been seen as a general approach which includes other sources of evidence. Messick [11] drew upon Loevinger's work to define construct validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other models of assessment" [11] p. 13. He argues that the traditional conception of validity is incomplete and fragmented because it lacks the value of consequences of outcome meaning and the public effects of the scores. Thus, he recommends that validity can be separated into the six aspects supported by evidence: content, substantive, structural, generalizability, external, and consequential aspects [11]. Messick's unified conceptualisation of construct validity [11], [22] offers a way of classifying the types of indications required to assist validity reasonings about the interpretation and utility of instruments to measure an intended construct.

Although Messick's validity framework has been considered by some to be difficult to interpret and apply [16], its model is still beneficial to individuals or organisations that either use information from assessments or are affected by the outcomes of the assessments' purpose. As a result, although different frameworks have been proposed and employed for the purpose of validation, Messick's [11], [22] central concepts for validity currently remain to be the most theoretically influential.

3.3 The Standard of Construct Validity Evidence

Hitherto, one of the most influential validity frameworks is in The Standards, which has been devised and promoted by the American Education Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) [1], [2]. This framework was first introduced in 1954 with an effort to introduce a standardised vocabulary and classification system for validity. AERA, APA, and NCME collaboratively published the Technical Recommendations for Psychological Tests and Diagnostic Techniques (later called The Standards). It was then modified twice in 1999 and 2014 with updated conceptions of validity. This framework presents the gold standard in validity in measurement in different nations and the United States of America, particularly. According to The Standards, validity is "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests", and that "the process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations" [1], [2], pp. 11. Five different strands of validity are manifested in this framework: test content, response processes, internal structure, relations to other variables, and consequential validity.

It should be pointed out that the notion of validity in this framework is presented as a single concept called construct validity. That means validity is considered an integral concept and the validation process could be regarded as establishing an integrated validity argument. This notion of validity comprising Messick's [22] definition has been impacted by many researchers.

In short, across the different validity frameworks, there are five common sources of validity which emphasizes the distinct elements that are necessary to be examined when assessing the validity of an instrument for a specific usage. Specifically, the most up to date validity framework is The Standards [1], [2]. This framework has been accepted and employed widely in literature [27], [19], [11], [22], [30], [7]. This review provides the foundation for the validation argument in the current study.

4. THE SECONDARY PISA CBA 2012 DATA

This study used previously collected data from the Programme for International Student Assessment (PISA) Computer-Based Assessments (CBA) 2012 instruments. The secondary data, including released items and analysis reports from PISA CBA 2012, were downloaded from the OECD website (www.oecd.org/pisa/pisaproducts/database-cbapisa2012.htm). Data from the participants of 32 countries/economies, with 117,933 students, were used in this study. The released computer-based items can be seen at www.oecd.org/pisa or http://erasq.acer.edu.au. No new reading material was released after the 2012 survey administration. There were 30 released items including both multiple-choice and partial-credit formats. There are two reasons why this study focused only on the computer-based PISA 2012 items. First, for an equal comparison between constructs, the method of delivering the tests needs to be the same in the three assessments. The interactive PS assessment was administered in a computer environment. As a result, the computer-based mathematics test and the digital reading test in PISA 2012 were chosen as having a similar administered approach. There were also a science test and a financial test in the PISA 2012 cycles; however, they were only administered in paper-and-pencil form. Thus, neither was included in this research. Second, a requirement of this study was that each student had to answer all three assessments at a similar time. Thus, only the PISA 2012 CBA data could satisfy this condition. Notable insights generated in this section also be found in Nguyen et al. [31].

5. RESULTS AND DISCUSSION

As discussed in previous sections concerning the literature consulted for an appropriate methodology, on account of its broad influence and clear guidelines, the Standards [1], [2] have been adopted to guide the exploration of the existence of the Generic PS construct underlying both the domain-specific and domain-general PS constructs in this study. Additionally, the set of components and tools for a validity argument model by Chapelle, Enright, and Jamieson [9] is employed to provide further clarification about the stages taken in this project to make a validity argument. It should be noted that the stages are iterative rather than sequential and lay the foundation for validity exploration.

The VE model illustrated in Figure 1 represents an elaborate enterprise and a serial, progressive procedure aligned with the content and structural validity aspects of The Standards [1], [2] framework. Other validity aspects were beyond the scope of this study and will be explored in the future. The procedure was formulated by a validity exploration planned to establish to

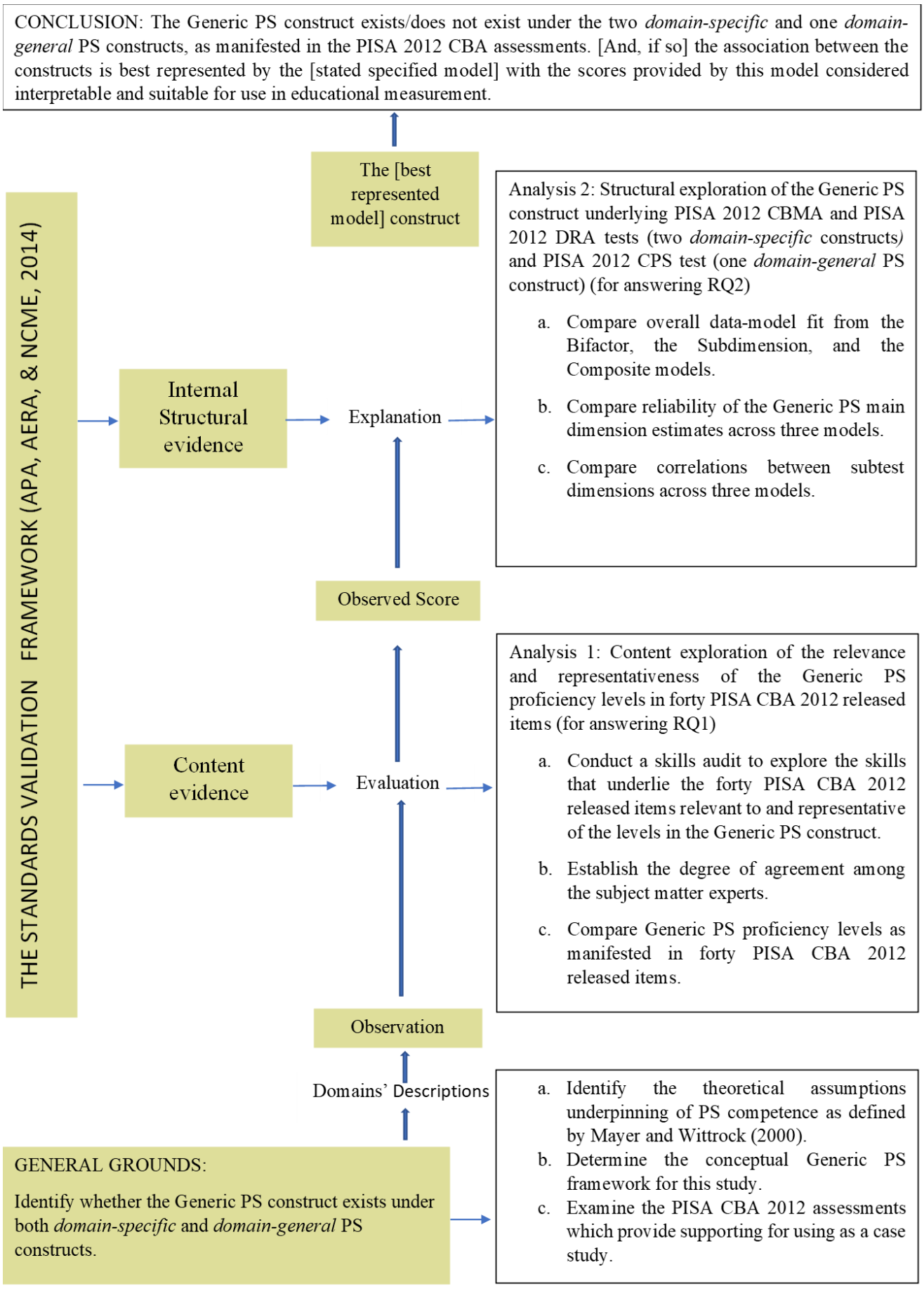

Figure 1. Validity exploration model related to the existence of a generic PS construct underlying domain-specific and domain-general PS constructs (modified and adapted from Chapelle et al. [9]).

what degree a Generic PS construct exists underlying both domain-specific and domain-general PS constructs. The examinations and related claims rest on one another to answer the general research question. The arrows indicate that the assumptions and assisting proof are orientated with every phase shaping the groundwork for the next exploration to offer sufficient constructive proof to endorse the existence of a Generic PS construct underlying the domain-specific and domain-general PS constructs.

The validity exploration schematic is read starting at the bottom of the diagram. In this instance, the question is whether or not the Generic PS construct exists underlying domain-specific and domain-general PS constructs. The general theoretical grounding was based on (i) Mayer and Wittrock's [21] definition of PS competence, (ii) the literature review about Generic competence, and (iii) an examination of the existence of related PS terms in PISA 2012 CBA domains' description [31]. The finding that a general conceptualisation of students' performance in Generic PS underlies both domain-specific and domain-general PS constructs as manifested in PISA 2012 CBA tasks serves as a foundation for the next step of inference, evaluation.

In Analysis 1, at the first stage, the question is, to what degree are the competencies which underlie the 40 PISA 2012 CBA released items relevant to and representative of the proficiency levels in the Generic PS scale. Notable insights generated in this analysis can be found in Nguyen et al. [32]. Then, the Generic PS skills in both the domain-specific and domain-general PS constructs are compared to reveal the existence of the Generic PS proficiency levels.

In Analysis 2, the observed score is the grounds for the explanation of inference, which links the observed score to the model of the association between the Generic PS construct and three PS constructs. This exploration of internal structural validity was undertaken as part of Analysis 2. It was supported through several forms and stages of analysis typically associated with seeking evidence based on a scale's internal structure. Three Rasch-based models were specified under the IRT framework: the Bifactor, Subdimensional, and Composite models. These members of the Rasch models are increasingly employed in analysing cognitive developmental data [33], [34]. These models were compared in terms of the fit indices, the IRT model subscale correlations, and the reliability estimates.

Finally, after providing structural explorations of the relationships between the different constructs, considerations are made about the usefulness of the best representative model for educational purposes. Together, the results provide an integrated exploration of the existence of the Generic PS competence underlying the two domain-specific (as manifested in PISA 2012 CBMA and PISA 2012 DRA) and one domain-general PS constructs (as manifested in PISA 2012 CPS).

6. CONCLUSION, IMPLICATION AND RECOMMENDATIONS

This paper has reviewed previous approaches for assessing the existence of a generic PS construct across domain-general and domain-specific PS constructs and discussed the shortcomings of these approaches referring to validity and reliability. The information obtained is critical to understand and evaluate how the exploration of the existence of generic PS across these PS constructs has been researched or investigated through different approaches.

It was believed that the use of such a range of data analysis methods would enhance the chance to investigate the existence of a generic PS construct across domain-general and domain-specific PS constructs, as manifested in PISA CBA 2012. Thus, the VE model is presented in this paper. The fundamental details, specific techniques, and procedures for evidence collection and analyses are presented in the previous paper [26] and the next coming paper.

The VE model proposed in this study provides a useful way forward for the investigation of both sub-dimensions and a main generic dimension. The adoption of The Standards [1], [2] validation framework, as applied in the current research, provides a blueprint for other studies where researchers endeavour to find more evidence for the existence of a generic construct along with subtests' PS constructs. The VE model proposed in this study could be adapted in different learning areas such as in physics and language education, or any other fields with multiple theoretically conceived main domains and sub-domains.

Besides, further studies could be carried out to examine additional validation exploration that may give more support for the existence of generic PS construct across the three PS constructs. This work could involve an examination of external, consequential, and generalizability aspects. The focus of the research was to evaluate the existence of the Generic PS construct across different learning domains based on the OECD sample design for 15-year-old students from 32 countries. Exploration of the differences across countries or gender groups may be fruitful areas for further work.

REFERENCES

[1] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education: Standards for educational and psychological testing, (1999), Washington, DC: American Educational Research Association.

[2] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education: Standards for educational and psychological testing, (2014). (0935302255), Washington, DC: American Educational Research Association.

[3] A. Fischer, S. Greiff, S. Wüstenberg, J. Fleischer, F. Buchwald, J. Funke, Assessing analytic and interactive aspects of problem solving competency, Learn. Individ. Differ. 9, (2015), pp. 172–9. DOI: 10.1016/j.lindif.2015.02.008

[4] J. Raven, Psychometrics, cognitive ability, and occupational performance, Review of Psychology, 7(1-2), (2000), pp. 51-74.

[5] L. L. Baird, Review of problem solving skills, ETS Research Report Series 1, (1983).

[6] R. Scherer, R. Tiemann, Evidence on the effects of task interactivity and grade level on thinking skills involved in complex problem solving, Thinking Skills and Creativity, 11, (2014), pp. 48-64.
DOI: 10.1016/j.tsc.2013.10.003

[7] J. Wirth, E. Klieme, Computer-based assessment of problem solving competence, Assessment in Education: Principles, Policy & Practice, 10(3), (2003), pp. 329-345.
DOI: 10.1080/0969594032000148172

[8] G. Molnár, S. Greiff, B. Csapó, Inductive reasoning, domain specific and complex problem solving: Relations and development, Thinking Skills and Creativity, 9, (2013), pp. 35-45.
DOI: 10.1016/j.tsc.2013.03.002

[9] C. A. Chapelle, M. K. Enright, J. Jamieson, Does an argument-based approach to validity make a difference? Educational Measurement: Issues and Practice, 29(1), (2010), pp. 3-13.
DOI: 10.1111/j.1745-3992.2009.00165.x

[10] J. C. Alderson, Language testing in the 1990s: How far have we come? How much further have we to go? In S. Anivan (Ed.), Current developments in languages testing, Vol. 25, (1991), pp. 1-26, Singapore: SEAMEO Regional Language Centre.
[11] S. Messick, Validity, in R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-103). (New York: American Council on Education), 3rd ed., (1989), pp. 13-103.
[12] E. L. Thorndike, The nature, purposes, and general methods of measurements of educational products, in National society for the study of education, 17th yearbook of the national society for the study of education, part II Bloomington, IL: Public School Publishing Co, (1918), pp. 16-24.
[13] G. Molnár, S. Greiff, S. Wüstenberg, A. Fischer, Empirical study of computer-based assessment of domain-general complex problem-solving skills, in B. Csapó & J. Funke (Eds.), The nature of problem solving: Using resarch to inspire 21st century learning, Paris: OECD Publishing, (2017), pp. 125-140.
DOI: 10.1787/9789264273955-10-en
[14] D. C. Briggs, M. Wilson, An introduction to multidimensional measurement using Rasch models, Journal of Applied Measurement, 4(1), (2003), pp. 87-100.
[15] R. J. Adams, Reliability as a measurement design effect, Studies in Educational Evaluation, 31(2) (2005), pp. 162-172.
DOI: 10.1016/j.stueduc.2005.05.008
[16] M. Wu, H. P. Tam, T. Jen, Educational measurement for applied researchers: Theory into practice. Singapore: Springer (2016).
[17] R. Scherer, R. Tiemann, Factors of problem-solving competency in a virtual chemistry environment: The role of metacognitive knowledge about strategies, Computers & Education, 59(4), (2012), pp. 1199-1214.
DOI: 10.1016/j.compedu.2012.05.020
[18] P. De Boeck, M. Wilson, Explanatory item response models: a generalized linear and nonlinear approach, Springer, (2004).
[19] M. T. Kane, So much remains the same: Conception and status of validation in setting standards, In Setting performance standards: Concepts, methods, and perspectives, Mahwah, NJ: Lawrence Erlbaum Associates, (2001) pp. 53-88.
[20] E. E. Cureton, Validity, In E. F. Lindquist (Ed.), Educational measurement (1951), pp. 621-694, Washington, DC: American Council on Education.
[21] R. E. Mayer, M. C. Wittrock, Problem solving, In P. A. Alexander & P. H. Winne (Eds.), Handbook of educational psychology, New York: Macmillan, Vol. 2, (2006), pp. 287-303.
[22] S. Messick, Validity of test interpretation and use, in M. C. Alkin (Ed.), Encyclopedia of educational research, New York: Macmillan, (1992), pp. 1487-1495.
[23] A. Anastasi, Evolving concepts of test validation, Annual Reviews Psychology, 37(1), (1986), pp. 1-16.
DOI: 10.1146/annurev.ps.37.020186.000245
[24] T. L. Kelley, Interpretation of educational measurements, New York: MacMillan, (1927).
[25] W. H. Angoff,. Validity: An evolving concept. In H. Wainer & H. I. Braun (Eds.), Test validity (1988), pp. 19-32. Hillsdale, NJ: Lawrence Erlbaum.
[26] D. Borsboom, G. J. Mellenbergh, J. Van Heerden, The concept of validity, Psychological Review, 111(4), (2004), p. 1061.
DOI: 10.1037/0033-295X.111.4.1061
[27] L. J. Cronbach, P. E. Meehl, Construct validity in psychological tests. Psychological Bulletin, 52(4), (1955), p. 281.
DOI: 10.1037/h0040957
[28] J. Loevinger, Objective tests as instruments of psychological theory. Psychological Reports, 3(3), 91957), pp. 635-694. DOI: 10.2466/pr0.1957.3.3.635
[29] H. Eklöf, Motivational beliefs in the TIMSS 2003 context: Theory, measurement and relation to test performance (2006), Doctoral dissertation, Institutionen för beteendevetskapliga mätningar, Umeå universitet, Sweden.
[30] E. W. Wolfe, J. E. Smith, Instrument development tools and activities for measure validation using Rasch models: Part II--validation activities, Journal of Applied Measurement, 8(2), (2007), pp. 204-234.
[31] Organisation for Economic Co-operation and Development (OECD), PISA 2012 results: What students know and can do: Student performance in mathematics, reading and science, Paris: OECD, (2014).
[32] L. A. Nguyen Khoa, T. K. C. Nguyen, R. J. Adams, Assessment of the generic problem-solving construct across different contexts, Journal of Physics: Conference Series, 1379(1), 012059, (2019).
[33] T. L. Dawson, Y. Xie, M. Wilson, Domain-general and domain-specific developmental assessments: do they measure the same thing? Cognitive Development, 18(1), (2003), pp. 61-78.
DOI: 10.1016/S0885-2014(02)00162-4
[34] T. L. Dawson, A stage is a stage is a stage: A direct comparison of two scoring systems, The Journal of Genetic Psychology, 164(3), (2003), pp.335-364.
DOI: 10.1080/00221320309597987