On Demonstrating the Bifactor Model with 2015 PISA Collaborative Problem-Solving Items

Lingde Kong

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Education

University of Washington

2023

Committee:

Chun Wang

Elizabeth Sanders

Program Authorized to Offer Degree:

Education

University of Washington

Abstract

On Demonstrating the Bifactor Model with 2015 PISA Collaborative Problem-Solving Items

Lingde Kong

Chair of the Supervisory Committee:

Chun Wang

Measurement and Statistics

New assessments built on theoretical frameworks require effective measurement tools to evaluate their psychometric properties. However, conventional unidimensional model assumes that the latent construct is unidimensional, without accounting for interdependencies among sub-factors. In contrast, the bifactor model offers an alternative framework that incorporates a general factor and orthogonal skill-specific factors, enabling a comprehensive evaluation of constructs with underlying subfactors. This project demonstrates use of the bifactor model in evaluating assessments for constructs with multiple dimensions, with a focus on the collaborative problem-solving (CPS) competency assessment items from the 2015 Programme for International Student Assessment (PISA). The study focuses specifically on 15-year-old students from four Chinese provinces in Mainland China, with a sample size of N = 1,675 participants. EFA determines the optimal number of factors, followed by a comprehensive CFA using empirically driven and theory driven models. Findings demonstrate that the empirically driven bifactor model outperforms other theory driven MIRT models, exhibiting the best model fit. The

implications identify problematic items requiring further investigation. This paper contributes to the relatively unexplored realm of CPS assessment design using the bifactor model as an analytic tool. The analysis and findings contribute to the scholarly understanding and offer practical implications for practitioners and researchers in the field.

Keywords: bifactor model, collaborative problem-solving, Multidimensional Item Response Theory, MIRT, PISA

On Demonstrating the Bifactor Model with 2015 PISA Collaborative Problem-Solving Items

It is not uncommon for contemporary assessments to be constructed based on a multidimensional framework. Considerations of dimensionality play a vital role in the development and ongoing validation for those encompassing skill-specific factors (Edelen et al., 2007; Dunn et al., 2020). This is a critical issue, particularly when evaluating newly developed assessments, as such analyses provide researchers with valuable insights into whether the scale effectively captures the underlying theoretical framework upon which it is built. Moreover, an understanding of dimensionality influences practical decisions regarding score reporting. However, for constructs involving skill-specific grouping factors, there is no widespread consensus on how dimensionality should be assessed, or which model should be employed.

The objective of this master's thesis is to demonstrate use of the bifactor item response theory model, along with other multidimensional item response theory (MIRT) models, as a tool to shed light on the factor structure of constructs that feature plausible subscales. The contextual focus of this study lies on the collaborative problem-solving (CPS) competency assessment from the 2015 Programme for International Student Assessment (PISA).

The paper initially conducts exploratory factor analyses (EFAs) to determine the optimal number of factors and factor loading structure, followed by confirmatory factor analyses (CFAs) within an item response theory (IRT) framework to compare various CFA model structures, including the bifactor model and the between-factor model. This analysis is conducted to examine both the theoretical implications and practical considerations relevant to the audience interested in research of Collaborative Problem-Solving (CPS). Additionally, the thesis investigates the incorporation of testlets to gain insights into the independence assumption of the

assessment design. The series of MIRT analyses will also provide valuable insights for the refinement of the assessment items. The paper intends to promote the adoption of the bifactor model as one of the measurement models for CPS research and for other constructs with skill-specific subdimensions.

Current Research on Collaborative Problem Solving

To gain a comprehensive understanding of the dimensionality challenges associated with constructs theoretically designed to encompass multiple skill-specific dimensions, this study is contextually centered around the analysis of Collaborative Problem-solving (CPS) data from the 2015 Programme for International Student Assessment (PISA). The primary objective of this research is to investigate the factor structure of collaborative problem-solving, a critical skill set in both educational and occupational contexts.

Despite its recognized significance, the assessment of individual students' ability to effectively collaborate in problem-solving scenarios remains an under-researched area (OECD, 2017b). To address this research gap, the Programme for International Student Assessment (PISA) has defined collaborative problem-solving (CPS) competency as the capacity of an individual to engage in a process wherein two or more agents collaborate to solve a problem. This collaborative endeavor entails the sharing of understanding, effort, knowledge, skills, and resources towards reaching a solution. To measure individual-level CPS skills, PISA has developed a computer-based assessment, which focuses on 12 weighted collaborative problem-solving skills (refer to Table 1). These skills emerge as a fusion of four distinct individual problem-solving processes (A-D) and three collaborative competencies (1-3).

Multidimensional IRT models

In the field of psychometrics, researchers have developed various measurement models to capture the complexities of constructs with multiple dimensions. The application of Item Response Theory (IRT) has revolutionized the measurement of latent constructs, such as abilities, traits, or attitudes (Reckase, 2009; van der Linden, 2005). IRT models provide a robust framework for understanding the relationship between individuals' responses to test items and the latent construct being measured. While traditional unidimensional IRT models have been widely used, there has been growing recognition of the need to incorporate multidimensionality into the measurement models.

Unidimensional models, like traditional factor analysis, assume that observed variables reflect a single underlying factor (Reise et al, 2010). However, these models may not fully account for the multidimensionality often present in constructs.

To address this limitation, bi-factor models have been proposed. Bi-factor models include a general factor representing common variance across all indicators, along with specific factors capturing unique variance associated with each dimension or subdomain (Chen, West, & Sousa, 2006). These models allow us to examine the contributions of specific factors while controlling for the influence of the general factor.

Another approach is the between-item multidimensional item response theory (MIRT) models. These models consider multidimensionality by allowing item parameters to vary across different dimensions (Reckase, 2009). By estimating separate item parameters for each dimension, these models can better capture the multidimensional nature of the construct and provide more accurate measurements.

Furthermore, high-order (or second-order) multidimensional IRT (MIRT) models incorporate a specific dimension for each testlet, just like the bi-factor and the testlet models. It

also contains a general dimension, but, unlike in the bi-factor models, items do not directly depend on this general dimension. Rather, items only directly depend on their respective specific dimensions, which in turn depend on the general dimension. It is assumed that the specific dimensions are conditionally independent on the general factor. That is, all associations between the specific dimensions are assumed to be taken into account by the general dimension (Rijmen, F., 2010).

Testlet models are another type of constrained bifactor model that account for dependencies among items within a testlet. A testlet refers to a group of items sharing a common stimulus or context (Wainer & Kiely, 1987). Testlet models assume that the general factor influences performance at the testlet level, whereas specific factors operate at the item level within the testlet (Reise et al., 2010). It is a special case of the bi-factor model. It is obtained by constraining the loadings on the specific dimension to be proportional to the loadings on the general dimension within each testlet (Rijmen, 2010). Furthermore, the second-order model is formally equivalent to the testlet model by algebraic manipulations.

In summary, these measurement models offer valuable tools for understanding the complex nature of constructs with multiple dimensions. They provide researchers with flexibility in capturing multidimensional structures and enable more accurate assessment and interpretation of test scores. Since testlet model is a special case of the bifactor model with constraints on the loadings on the specific dimension to be proportional to the loadings on the general dimension within each test, and since the higher-order model is formally equivalent to the testlet model, the project will only focus on the comparison between the bifactor model and the between-item models. Further, we focus on the bifactor model for the CPS construct because it appears ideally suited for representing the construct-relevant multidimensionality that arises in the responses to

measures of broad constructs where multiple and distinct domains of item content are included to increase content validity (see, e.g., Reise, Moore, & Haviland, 2010).

Current Study

This aim of this study is to demonstrate use of the bifactor model in real research settings, within the context of collaborative problem-solving (CPS) as a construct. In doing so, it is hoped that the results of this research can make a valuable contribution to the relatively unexplored realm of the CPS assessment design. Specifically, the following research questions will be explored. The overarching question is: Does the CPS assessment effectively measure the collaboration dimensions from theoretical framework? Sub-questions include the following:

RQ1: What is the factor structure of Collaborative Problem Solving?

RQ2: Which factor structure would be most appropriate to capture the latent construct?

RQ3: Should testlets and empirically-driven between-item model be considered?

Method

Data Source

The data for this study is drawn from the Programme for International Student Assessment (PISA) 2015, which conducted a large-scale computer-based assessment of collaborative problem-solving competency. The release data consisted of 12 items that captured the cross-section between 4 individual problem-solving processes and 3 collaborative competencies. In this study, the three subdimensions of collaborative competencies, namely "Establishing and maintaining shared understanding" (SU), "Taking appropriate action" (AA), and "Establishing and maintaining team organization" (TO), will be adopted as skill-specific factors within the theory-inspired modeling framework. This choice is justified based on several

considerations. Firstly, the other four individual problem-solving processes primarily focus on the sequential order of solving problems, rather than directly addressing the competency itself. Consequently, these four subdimensions fail to adequately reflect the narrow subdomain constructs of the overall CPS (Collaborative Problem Solving) construct. Secondly, these subdimensions failed to capture the crucial element of collaboration within problem-solving tasks. Collaborative competencies are an essential aspect of CPS, encompassing the skills required to effectively work as a team and interact with others to achieve shared problem-solving goals. By contrast, the three identified collaborative competencies—SU, AA, and TO—are more aligned with the specific skills necessary for successful collaboration. These competencies provide a framework for considering the additional dependencies between items belonging to the same "collaboration" subdimension in relation to the general dimension of CPS.

PISA assessed approximately 15-year-old students from 56 countries and economies. For this research, the focus is on students from four PISA-participating Chinese provinces of Beijing, Shanghai, Jiangsu, and Guangdong. The final sample size used in the analysis was $N = 1{,}675$ participants after removing missing data. There are 12 test items from the released unit “Xander” (task 1 includes 5 items; task 2 includes 3 items; tasks 3 and 4 have 2 items respectively) (see figure 1). The item responses of this unit are dichotomously coded according to item-coding rubrics. To ensure an appropriate analysis, the dataset is divided into two subsets: training data for the exploratory factor analysis (EFA) with $N = 838$, and test data for the confirmatory factor analysis (CFA) also with $N = 837$.

Measures

The variables used in the analysis include the 12 multiple-choice items that assess collaborative problem-solving competency in the PISA 2015 assessment. All items were coded as dichotomous responses.

Data Analyses

Exploratory Factor Analysis (EFA) on Factor Numbers. The EFA was conducted to determine the appropriate number of factors in the factor structure. Models with 1 to 5 factors were examined and compared. The fit of the models was assessed based on several criteria, including the Akaike Information Criterion (AIC), the Sample-Adjusted BIC (SABIC), and chi-square tests. The factor loadings for each model were examined to understand the underlying factor structure (see Table 2).

Exploratory Factor Analysis (EFA) on MIRT Types. Following the initial EFA analysis, the focus shifted to examining the factor structure using a bifactor model. Based on theoretical considerations, both a 3-factor bifactor model and a 4-factor bifactor model were explored. The EFA with bifactor Q rotation was conducted, and the model fit was assessed. The factor loadings were examined to understand the structure of the collaborative problem-solving construct (see Table 3).

Confirmatory Factor Analysis (CFA) on comparing *empirically-driven and theory-driven bifactor Models.* In this step, two types of bifactor models were compared: an empirically-driven bifactor 4-factor model and a theory-driven bifactor 3-factor model. CFA was performed on these models, and the fit indices were examined to assess the goodness of fit. The factor loadings on the general factor and skill-specific factors were examined to determine the appropriateness of the models (see Table 4 and Table 5).

Confirmatory Factor Analysis (CFA) on investigating testlet-driven and empirical-based between-item models. To explore the inclusion of testlets and an empirical-based between-item model, two models were examined: the test-scenario-driven between-factor 3-factor model and the empirically-driven between-factor 3-factor model. CFA was conducted on these models, and the fit indices were examined to evaluate their adequacy. The findings were analyzed to determine the necessity of incorporating testlets and empirically-driven sub-factors (see Table 6 and Table 7).

Results

EFA Analysis - fit indices. The first research question is addressed by EFA analysis. Table 2 presents the fit indices of -2LL, AIC, BIC, and SABIC, as well as the chi-squared value. The result indicates that the 3-factor model yields the best fit (with evidence from AIC, SABIC, and chi-squares). However, taking theoretical framework of collaboration into consideration, the 4-factor EFA will also be considered (with 1 general factor and 3 skill-specific factors representing different aspects of collaboration).

EFA Analysis - loadings. Further, I compared the 3-factor EFA (rotation = oblimin) and 4-factor EFA (rotation = bifactor Q) (see Table 3). Based on the factor loadings, the 3-factor EFA and 4-factor bifactor EFA yield similar model fit; additionally, item 10 has small loadings in both situations. Therefore, item 10 might be removed in either situation for the empirical-based modelling in the following steps.

CFA Analysis - Empirically-driven bifactor 4-factor model. To address the second research question in terms of the best MIRT model for capturing the underlying structure of the CPS construct, various MIRT models based on empirical driven and theory-driven frameworks were employed. First, driven by the empirical evidence, I constrained the loadings on each item

based on the empirical results in the CFA analysis (rotation = oblimin) with item 10 removed. As the theoretical framework encompasses three skill-specific factors, the loadings derived from the 4-factor CFA were utilized, assuming one factor as the general factor and the remaining three factors as hypothetical skill-specific factors. Building on the empirical 4-factor CFA result, item 7 is only significantly loaded onto the general factor and has minimal loadings on all other factors, therefore, the item 7 is constrained to be only loaded onto by the general factor. In this empirical-driven model, items 1 and 9 did not exhibit significant loadings on any factor. Furthermore, it was observed that four items exhibited significant loadings on the first skill-specific factor, which corresponds to the SU factor. This finding suggests that these items were appropriately designed to capture and predict the underlying construct of the SU factor. However, only one item loaded onto the second skill-specific factor, and no items loaded onto the third factor (see Table 4 and Figure 2). Consequently, the overall significance of loadings was less than optimal, suggesting further consideration of a 3-factor bifactor model.

CFA Analysis - Theory- driven bifactor 3-factor model. Subsequently, a theory-based model was conducted by imposing loadings constraints in accordance with the theoretical framework. For instance, items designed to measure the SU factor (denoted by the _SU suffix) were constrained to load onto both the general factor and the first factor. Similarly, items designed for the TO factor (denoted by the _TO suffix) were constrained to load onto both the general factor and the second factor. Given that only one item loaded onto the AA (Taking appropriate action) factor, it was exclusively set to load onto the general factor (see Figure 3). Consequently, this model represents a 3-factor structure consisting of one general factor and two theoretical skill-specific factors (see Table 5 and Figure 3). In this theory-driven bifactor 3-factor model, item 8 did not load onto any factors, raising concerns about its suitability for the

assessment goal. Additionally, items 1, 5, 7, 8, and 10 did not yield significant loadings on their respective skill-specific factors. This indicates that a 3-factor theory-based bifactor model may not adequately represent the theoretical framework. Therefore, a 3-factor between-item model will be further explored.

CFA Analysis - Test-scenario between-item 3-factor model. To examine if the items in the same unit are dependent due to shared context, the test-scenario between-item 3-factor model was applied (see Table 6 and Figure 4). The factor loadings for all 5 items in Part I are significant; 2 out of 3 items in Part II are significant; 3 out of 4 items in Part III are significant. This indicates that there is a potential violation of the IRT assumption of local independence due to the context-dependent items (DeMars, 2006).

CFA Analysis - Empirical-driven between-item 3-factor model. Lastly, the analysis on the empirical-driven 3-factor model in a between-item framework was conducted to see if the general-factor assumed in the bifactor model is necessary (See Table 7 and Figure 5). The result reveals a decent amount of significant loadings with only item 1 and item 9 not significantly loaded onto any factors.

CFA Analysis – fit Indices. Finally, the fit indices of the models discussed above showed that the empirical-based 4-factor bifactor model yields the best fit even though the loadings does not seem to be ideal enough (See Table 8). This indicates that overall, the model could capture the theoretical framework of CPS where it is centered around; but several items should be investigated further to ensure their predictability on the general factors as well as the skill-specific factors that it is designed for.

Discussion

The present study's major findings are two-fold. First, empirical-driven bifactor models provide support for the three skill-specific factors outlined in the theoretical framework of collaborative problem-solving (CPS). This indicates that the assessment effectively captures the intended dimensions of CPS. Additionally, the bifactor model offers insights into problematic items by examining the loadings on both the general factor and the skill-specific factors. It serves as an alternative modeling framework that is particularly valuable for evaluating the empirical plausibility of subscales within the CPS assessment.

Limitations

The study also acknowledges several limitations. First, the participants included in the analysis may not be fully representative of the larger population. Second, missing data were not considered in the analysis, which could potentially impact the results. Third, there is a potential violation of the assumption of local independence in the item response theory (IRT) due to the context-dependent items. This should be taken into consideration when evaluating the assessment based on empirical-based or theoretical-based tests. Lastly, the sampling weight was not considered in this stage of the analysis.

Conclusion

The use of the bifactor model has important implications for capturing the collaborative problem-solving construct and other constructs that exhibit categorical sub-factors along with a general factor based on theoretical frameworks. It provides a valuable tool for effectively representing the multidimensional nature of these constructs.

For future research, it would be beneficial to explore the practical application of combining the bifactor model approach with testlet modeling. This is particularly relevant due to the potential violation of the assumption of local independence caused by context-dependent

items. Additionally, further investigation could focus on collecting more evidence regarding the necessity of reporting subscores for the different skill-specific abilities within the CPS construct. If distinct aspects of ability are identified, reporting subscores separately could enhance the assessment's capacity to capture these specific skills.

References

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research, 41*(2), 189-225.

Dunn, K. J., & McCray, G. (2020). The place of the Bifactor model in confirmatory factor analysis investigations into construct dimensionality in language testing. Frontiers in Psychology, 11(1357), 1– 16. https://doi.org/10.3389/fpsyg.2020.01357

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Qual. Life Res. 16*(Suppl. 1) 5–18. doi: 10.1007/s11136-007-9198-0

Häkkinen, P., Järvelä, S., Mäkitalo-Siegl, K., Ahonen, A., Näykki, P., & Valtonen, T. (2017). applications. Springer Science & Business Media.

Md Desa, Z. N. D. (2012). Bi-factor multidimensional item response theory modeling for subscore estimation, reliability, and classification (Doctoral dissertation, University of Kansas). Retrieved from http://kuscholarworks.ku.edu/dspace/handle/1808/10126

Preparing teacher-students for twenty-first-century learning practices (PREP 21): A framework for enhancing collaborative problem-solving and strategic learning skills. *Teachers and Teaching, 23*(1), 25–41.

Reckase, M. D. (2009). Multidimensional item response theory. Springer Science & Business Media.

Reise, S. P. (2012). The Rediscovery of Bifactor Measurement Models. *Multivariate Behavioral Research, 47*(5), 667-696. doi:10.1080/00273171.2012.715555

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment, 92*(6), 544-559.

Rijmen, F. (2010). Formal Relations and an Empirical Comparison among the Bi-Factor, the Testlet, and a Second-Order Multidimensional IRT Model. *Journal of Educational Measurement, 47*(3), 361-372.

van der Linden, W. J. (2005). Introduction to item response theory models and their applications, 2nd edition. New York: Springer.

Table 1

Matrix of CPS Construct with Allocation of Weights for PISA 2015

The 12-cell matrix illustrating the 12 CPS skills in the PISA 2015 assessment.

	(1) Establishing and maintaining shared understanding	(2) Taking appropriate action to solve the problem	(3) Establishing and maintaining team organisation
(A) Exploring and Understanding	(A1) Discovering perspectives and abilities of team members	(A2) Discovering the type of collaborative interaction to solve the problem, along with goals	(A3) Understanding roles to solve problem
(B) Representing and Formulating	(B1) Building a shared representation and negotiating the meaning of the problem (common ground)	(B2) Identifying and describing tasks to be completed	(B3) Describe roles and team organisation (communication protocol/rules of engagement)
(C) Planning and Executing	(C1) Communicating with team members about the actions to be/being performed	(C2) Enacting plans	(C3) Following rules of engagement, (e.g., prompting other team members to perform their tasks.)
(D) Monitoring and Reflecting	(D1) Monitoring and repairing the shared understanding	(D2) Monitoring results of actions and evaluating success in solving the problem	(D3) Monitoring, providing feedback and adapting the team organisation and roles

Note. Drawn from PISA 2015 collaborative problem-solving framework (p. 137), by OECD (2017) (https://www.oecd-ilibrary.org/education/pisa-2015-assessment-and-analytical-framework/pisa-2015-collaborative-problem-solving-framework_9789264281820-8-en;jsessionid=pz-w4V7Yiz1e8XwST5qmMwP_.ip-10-240-5-53).

Table 2

EFA Model Comparison for number of factors

Model	Parms	-2LL	AIC	BIC	SABIC	χ2 (df)	
M1: uni-dimensional model	24	-5698	11443	**11557**	11480		
M1.2: 2-factor EFA	35	-5675	11420	11585	**11474**	45.23 (11)	<0.001 ***
M1.3: 3-factor EFA	45	-5660	**11410**	11623	**11477**	33.08 (10)	**<0.001** ***
M1.4: 4-factor EFA	54	-5653	11414	11670	11498	10.44 (9)	0.316
M1.5: 5 factor EFA	62	**-5644**	11412	11706	11509	18.00 (8)	0.021 *

Note. *N* = 838 participants (split dataset randomly drawn from the original dataset with 1675 participants for EFA). Parms = number of paramaters estimated; LL = log-likelihood; CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized room mean residual. Change values compare the immediate prior model, except the first model.

* $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$.

Table 3

Exploratory Factor Analysis loadings - 3 factors and 4 factors

			Rotation: Oblimin			Rotation: bifactorQ			Rotation: bifactorQ			
Item	Description	Comm.	Factor1 Loading	Factor2 Loading	Factor3 Loading	Factor1 Loading	Factor2 Loading	Factor3 Loading	Factor1 Loading	Factor2 Loading	Factor3 Loading	Factor4 Loading
1	I1_G11_TO	0.08	**0.28**	0.03	-0.01	**0.19**	0.02	**0.23**	**0.22**	0.03	**0.07**	-0.21
2	I2_G12_SU	0.50	0.09	0.06	**0.65**	**0.71**	0.04	0.04	**0.61**	**0.23**	0.01	0.07
3	I3_G13_SU	0.22	-0.04	-0.04	**0.49**	**0.46**	-0.06	-0.06	**0.37**	**0.19**	-0.03	0.11
4	I4_G14_SU	0.52	-0.02	0.14	**0.71**	**0.71**	0.11	-0.05	**0.80**	0.00	-0.02	**0.46**
5	I5_G15_TO	0.32	**0.53**	0.04	0.04	**0.41**	0.04	**0.42**	**0.47**	**0.02**	-0.07	-0.26
6	I6_G21_SU	0.47	0.03	-0.22	**0.64**	**0.64**	-0.25	-0.01	**0.41**	**0.86**	0.01	0.00
7	I7_G22_TO	0.58	**0.77**	-0.01	-0.02	**0.51**	-0.01	**0.61**	**0.59**	-0.01	-0.01	-0.42
8	I8_G23_TO	0.83	0.01	**0.91**	0.01	0.10	**0.91**	0.00	**0.23**	-0.30	**0.22**	-0.02
9	I9_G31_TO	0.11	0.03	**0.24**	0.19	**0.23**	**0.23**	0.01	**0.27**	0.01	**0.87**	0.00
10	I10_G32_SU	0.03	0.01	**0.13**	0.10	0.12	**0.13**	0.00	0.13	0.00	-0.08	-0.02
11	I11_G41_AA	0.14	**0.28**	-0.02	0.15	**0.34**	-0.03	**0.21**	**0.37**	**0.03**	-0.10	-0.15
12	I12_G42_TO	0.14	**0.26**	-0.05	0.16	**0.33**	-0.06	**0.20**	**0.34**	**0.07**	-0.11	-0.15
Rotated SS Loadings			**1.11**	**0.98**	**1.68**	**2.39**	0.97	0.69	**2.32**	0.93	0.84	0.56

Note. *N* = 838 participants. Commu. = Communality. SU = Collaborative problem-solving competency 1: Establishing and maintaining shared understanding. AA = Collaborative problem-solving competency 2: Taking appropriate action C3(TO) solve the problem.TO = Collaborative problem-solving competency 3: Establishing and maintaining team organisation.

Table 4

CFA: empirical-driven bifactor 4-factor model

Items	Description	General Factor Loading		Factor 1 Loading		Factor 2 Loading		Factor 3 Loading
1	I1_G11_TO	0.13		--		-0.10		--
2	I2_G12_SU	0.58	*	0.55	**	--		--
3	I3_G13_SU	0.29	***	0.21	*	--		--
4	I4_G14_SU	0.53	***	0.50	**	--		--
5	I5_G15_TO	0.64	***	--		-0.26		--
6	I6_G21_SU	0.39	***	0.32	**	--		--
7	I7_G22_TO	0.71	***	--		--		--
8	I8_G23_TO	0.05	***	--		--		0.45
9	I9_G31_TO	0.43		--		--		0.46
10	I11_G41_AA	0.28	***	--		0.24	*	--
11	I12_G42_TO	0.35	***	--		0.62		--
Rotated SS Loadings:		2.16		0.70		0.53		0.41
R-squared		0.20		0.06		0.05		0.04

Note. N = 837 participants (split dataset randomly drawn from the original dataset with 1675 participants for CFA).

Table 5

CFA: theory-driven bifactor 3-factor model

Item	Description	General Factor Loading		Factor 1 Loading		Factor 2 Loading	
1	I1_G11_TO	0.15	*			-0.09	
2	I2_G12_SU	0.62	***	0.50	**	--	
3	I3_G13_SU	0.31	***	0.18	*	--	
4	I4_G14_SU	0.54	***	0.50	**	--	
5	I5_G15_TO	0.60	***	--		0.06	
6	I6_G21_SU	0.42	***	0.28	**	--	
7	I7_G22_TO	0.69	***	--		0.14	
8	I8_G23_TO	-0.03		--		0.84	
9	I9_G31_TO	0.40	***	--		0.28	*
10	I10_G32_SU	0.29	**	-0.14		--	
11	I11_G41_AA	0.29	***			--	
12	I12_G42_TO	0.28	***			0.13	*
Rotated SS Loadings:		2.16		0.70		0.53	
R-squared		0.20		0.06		0.05	

Note. N = 837 participants.

Table 6

CFA: test-driven between-factor 3-factor model

Item	Description	Factor 1 Loading	Factor 2 Loading	Factor 3 Loading
1	I1_G11_TO	0.16 **		--
2	I2_G12_SU	0.78 ***		--
3	I3_G13_SU	0.36 ***		--
4	I4_G14_SU	0.72 ***		--
5	I5_G15_TO	0.51 ***		--
6	I6_G21_SU	--	0.55 ***	--
7	I7_G22_TO	--	0.67 ***	--
8	I8_G23_TO	--	0.03	--
9	I9_G31_TO	--	--	0.13
10	I10_G32_SU	--	--	0.23 *
11	I11_G41_AA	--	--	0.42 ***
12	I12_G42_TO	--	--	0.62 **
Rotated SS Loadings:		2.21	.62	0.84
R-squared		0.18	.05	0.07

Note. $N = 837$ participants.

Table 7

CFA: empirical-driven between-factor 3-factor model

Item	Description	Factor 1 Loading		Factor 2 Loading		Factor 3 Loading	
1.00	I1_G11_TO	0.06		--		--	
2.00	I2_G12_SU	--		--		0.79	***
3.00	I3_G13_SU	--		--		0.36	***
4.00	I4_G14_SU	--		--		0.74	***
5.00	I5_G15_TO	0.49	***	--		--	
6.00	I6_G21_SU	--		--		0.49	***
7.00	I7_G22_TO	0.89	*	--		--	
8.00	I8_G23_TO	--		0.22	***	--	
9.00	I9_G31_TO	--		0.90		--	
10.00	I11_G41_AA	0.26	***	--		--	
11.00	I12_G42_TO	0.31	***	--		--	
Rotated SS Loadings:		2.21		.62		0.84	
R-squared		0.18		.05		0.07	

Note. $N = 837$ participants.

Table 8

CFA Model Comparison: theory-driven Versus EFA-inspired

Model	-2LL	AIC	BIC	SABIC	RMSEA	Chi-Sq (df)			
bif_emp_4fac	-5384	10831	10983	10881	0				
bif_theo_3fac	-5663	11395	11561	11449	0	N/A	(N/A)	N/A	N/A
bt_test_3fac	-5711	11474	11597	11515	0	N/A	(N/A)	N/A	N/A
bt_emp_3fac	-5454	10955	11069	10993	0	N/A	(N/A)	N/A	N/A

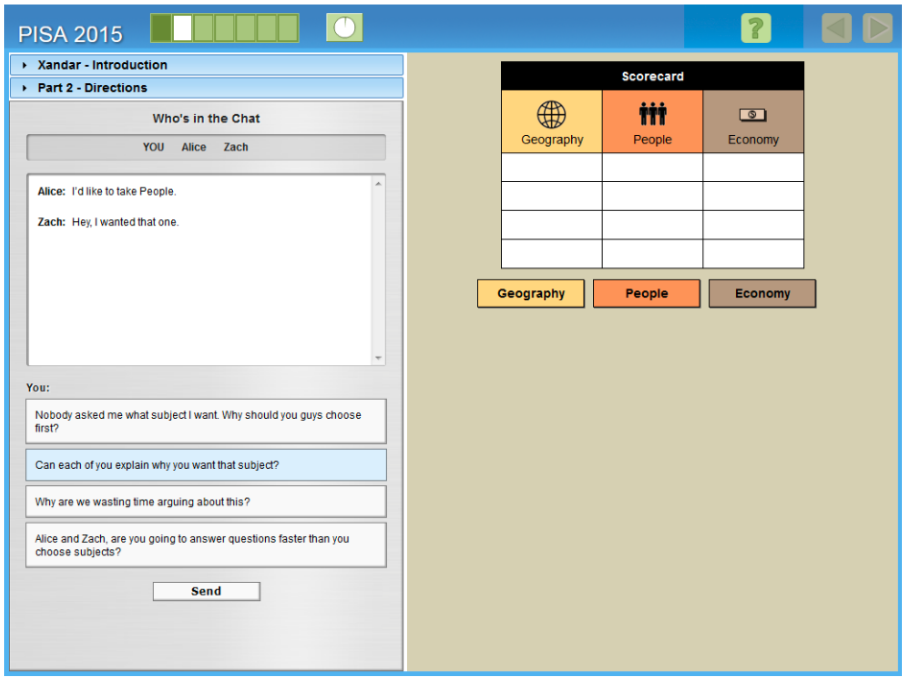
Note. *N* = 837 participants. LL = log-likelihood; CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized room mean residual. Change values compare the immediate prior model, except the first model; bif_emp_4fac = Empirical-driven bifactor 4-factor model; bif_theo_3fac = theoretical-driven bifactor 3-factor model; bt_test_3fac = Test-scenario between-item 3-factor model; bt_emp_3fac = Empirical-driven between-item 3-factor model.

Figure 1

Sample item from "Xander", the released unit of the CPS assessment

1.2.1. Part 2, Item 1: Choosing Subjects

Item	CC100201
Collaborative competency	Establishing and maintaining shared understanding 1
Problem-solving process	Exploring and understanding A
Collaborative problem-solving skill	Discovering perspectives and abilities of team members
Difficulty	598 (Level 3)
Credited response	"Can each of you explain why you want that subject?"



Note. Drawn from OECD (2015). Description of the released unit from the 2015 PISA collaborative problem-solving assessment, collaborative problem-solving skills, and proficiency levels. Retrieved November 23, 2021, from https://www.oecd.org/pisa/test/CPS-Xandar-scoring-guide.pdf

Figure 2

Factor loading diagram for confirmatory factor analysis (CFA) models fit: empirically-driven bifactor model (4-factor)

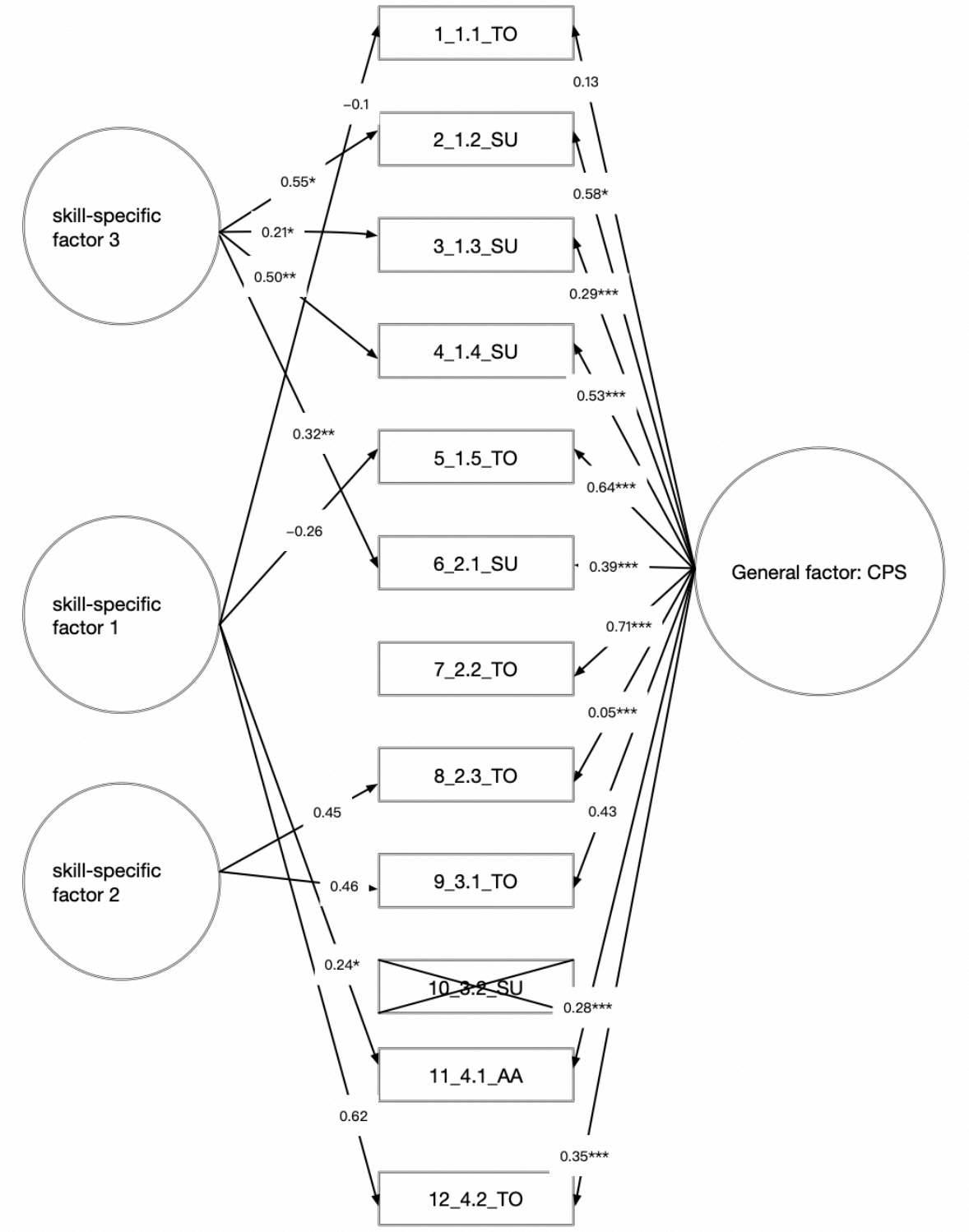


Figure 3

Factor loading diagram for confirmatory factor analysis (CFA) models fit: theoretically-driven bifactor model (3-factor)

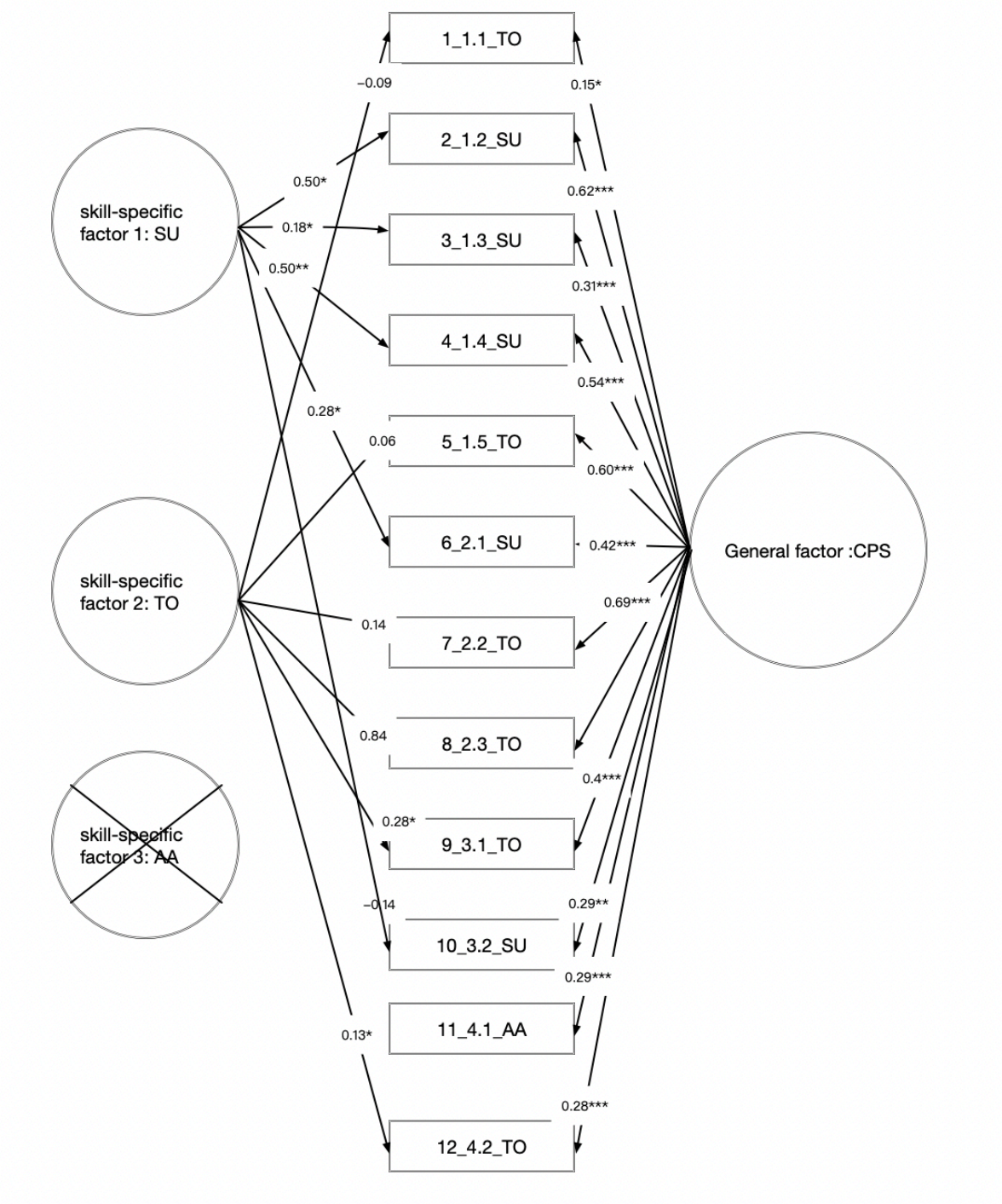


Figure 4

Factor loading diagram for confirmatory factor analysis (CFA) models fit: Test-scenario between-item model (3-factor)

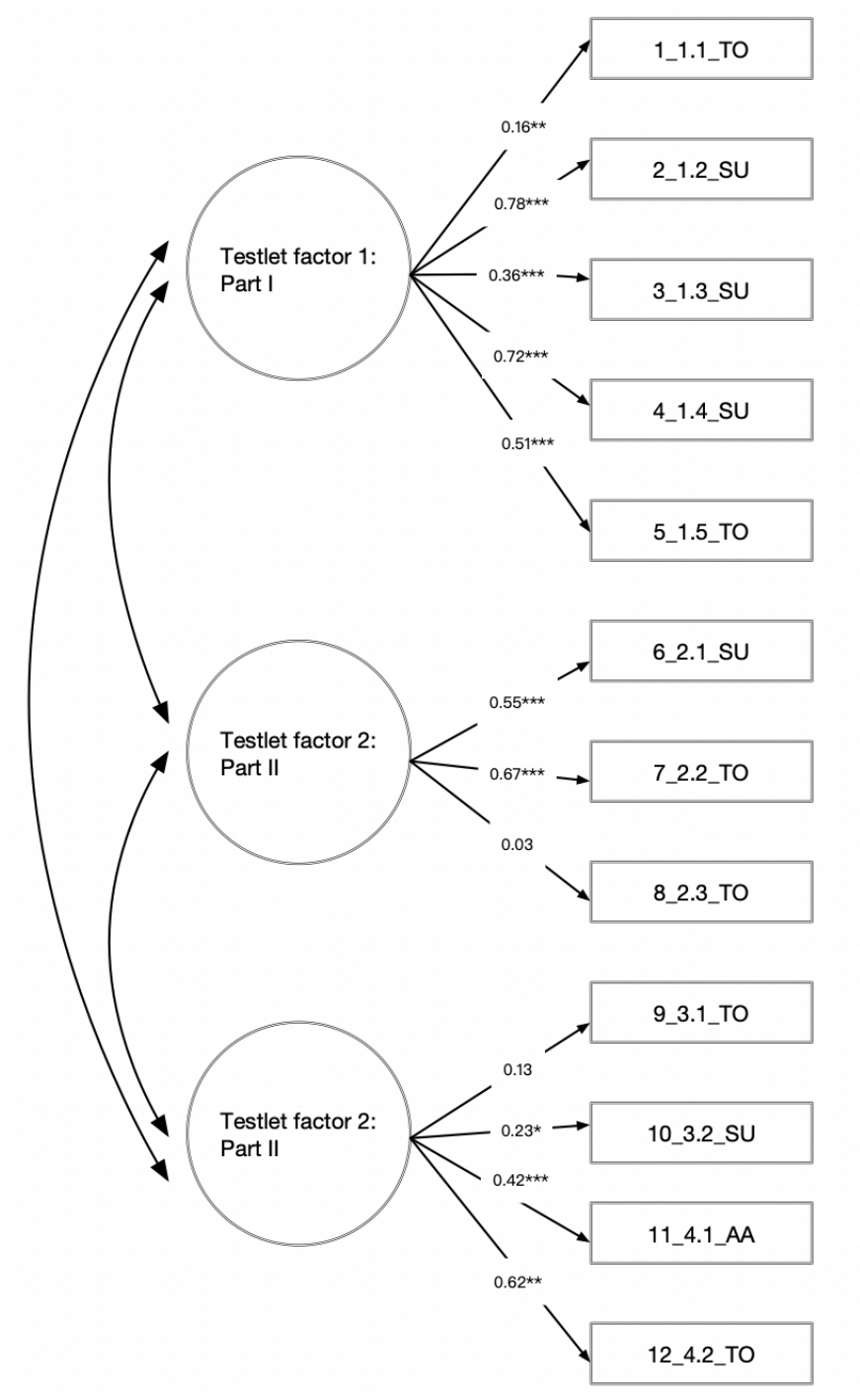
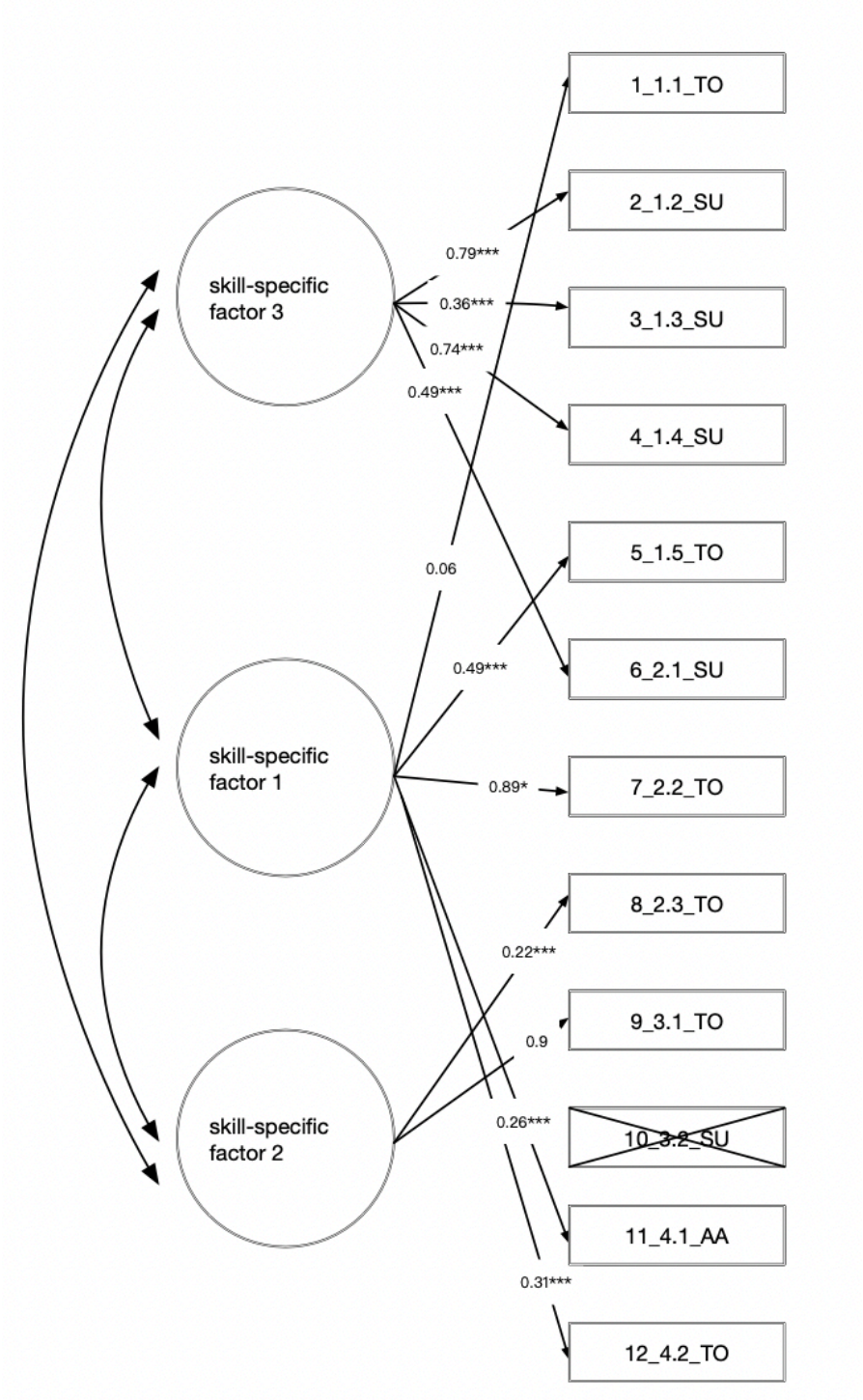


Figure 5

Factor loading diagram for confirmatory factor analysis (CFA) models fit: empirically-driven between-item model (3-factor)



Appendix A

Sample *R* Code

```
# unidimensional EFA
mod1n<-mirt(CPS_efa,1,SE=T)

# 2-factor EFA
mod1.2n<-mirt(CPS_efa,2,TOL=.001,SE=T,technical=list(NCYCLES=4000))

# 3-factor EFA
mod1.3n <-mirt(CPS_efa,3,TOL=.001,SE=T,technical=list(NCYCLES=20000))

# 4-factor EFA
mod1.4n<-mirt(CPS_efa,4,TOL=.001,SE=T,technical=list(NCYCLES=10000))

# 5-factor EFA
mod1.5n<-mirt(CPS_efa,5,TOL=.001,SE=T,technical=list(NCYCLES=10000))

# EFA model comparisons
anova(mod1n,mod1.2n)
anova(mod1.2n,mod1.3n)
anova(mod1.3n,mod1.4n)
anova(mod1.4n,mod1.5n)

# EFA loadings and SE
coef(mod1.3n,IRTpars = TRUE, printSE= T)
coef(mod1.4n,IRTpars = TRUE, printSE= T)


# CFA empirical-driven bifactor 4-factor model (table3: bif_emp_4fac)
specific <- c(2,1,1,1,2,1,NA,3,3,2,2)
bif_emp_4fac<- bfactor(data=CPS_cfa_no10,model = specific,TOL=.001,SE =
T,technical=list(NCYCLES=10000))

# CFA theoretical-driven bifactor 3-factor model (table4: bif_theo_3fac)
specific2 <- c(2,1,1,1,2,1,2,2,2,1,NA,2)
bif_theo_3fac <- bfactor(CPS_cfa,specific2,SE = TRUE,TOL=.001,technical=list(NCYCLES=10000))

# test-scenario-driven between-factor 30factor model(table5: bt_test_3fac)
test2<- 'F1 = 1-5
F2 = 6-8
F3 = 9-12
COV = F1*F2,F2*F3'
bt_test_3fac<-mirt(data=CPS_cfa, model=test2,TOL=.001,technical=list(NCYCLES=10000), SE=T)

# CFA empirical-driven between-factor 3-factor model (table 6: bt_emp_3fac)
test3<- 'F1 = 1,5,7,10-11
F2 = 8-9
F3 = 2-4,6
COV = F1*F2,F2*F3'
bt_emp_3fac<-mirt(data=CPS_cfa_no10, model=test3,TOL=.001,technical=list(NCYCLES=10000), SE=T)

# EFA model comparisons
anova(bif_emp_4fac,bif_theo_3fac)
anova(bif_emp_4fac,bt_test_3fac)
anova(bif_emp_4fac,bt_emp_3fac)

# CFA loadings and SE
coef(bif_emp_4fac, IRTpars = TRUE, printSE= T)
coef(bif_theo_3fac, IRTpars = TRUE, printSE= T)
coef(bt_test_3fac, IRTpars = TRUE, printSE= T)
coef(bt_emp_3fac, IRTpars = TRUE, printSE= T)
```