Supplementary Information

1. The reliability and validity of problem-solving skill scale with the spring test data

In this study, the measurement model of problem-solving skill was examined by convergent validity, internal consistency reliability and discriminant validity. Convergent validity is assessed with factor loadings, composite reliability (CR) and average variance extracted (AVE). As shown in Table S1, all factor loadings ranged from 0.701 to 0.918, exceeding the threshold value of 0.5; the CR ranged from 0.882 to 0.919, exceeding the threshold value of 0.7; and the AVE ranged from 0.599 to 0.674, exceeding the threshold value of 0.5. The Cronbach's alpha value exceeded the cut-off value of 0.7, which indicates that the internal consistency of the instrument was acceptable.

Table S2 shows that the discriminant validity was satisfactory because the square root of AVE of the four constructs exceeded the threshold value of 0.5 and was larger than the correlation values between each construct. Accordingly, the measurement model of problem-solving skill was acceptable.

Table S1 *Coefficients for the measurement model of Problem-solving skill*

Construct	Loadings	AVE	CR	Cronbach's alpha
Understanding (UD)		0.599	0.882	0.881
UD 1	0.800			
UD 2	0.753			
UD 3	0.798			
UD 4	0.701			
UD 5	0.813			
Representing (RP)		0.674	0.891	0.892
RP1	0.812			
RP2	0.832			
RP3	0.918			
RP4	0.710			
Executing (EC)		0.655	0.919	0.918
EC1	0.811			
EC2	0.751			
EC3	0.783			
EC4	0.833			
EC5	0.795			
EC6	0.878			
Reflecting (RF)		0.651	0.903	0.901
RF1	0.867			
RF2	0.713			
RF3	0.815			

RF4	0.738
RF5	0.886

Note. $N = 96$; AVE, average variance extracted; CR, composite reliability.

Table S2 *The correlations of each construct*

Construct	UD	RP	EC	RF
UD	0.774			
RP	-0.134	0.821		
EC	-0.029	0.628	0.81	
RF	-0.039	0.273	0.432	0.807

Note. UD, understanding; RP, representing; EC, executing; RF, reflecting.

2.The raw data output for the independent samples t-test for Table 2.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Initial	Equal variances assumed	.218	.642	.221	94	.826	.590	2.673	-4.718	5.899
	Equal variances not assumed			.220	92.464	.826	.590	2.678	-4.729	5.910
Fall	Equal variances assumed	6.661	.011	2.089	94	.039	7.742	3.706	.384	15.100
	Equal variances not assumed			2.122	86.454	.037	7.742	3.648	.491	14.993
Spring	Equal variances assumed	22.236	.000	2.985	94	.004	11.853	3.971	3.968	19.738
	Equal variances not assumed			3.064	72.064	.003	11.853	3.869	4.141	19.565

3.The raw data output for the independent samples t-test for Table 3.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
InitialPSS	Equal variances assumed	.087	.768	.697	94	.487	1.483	2.127	-2.741	5.706
	Equal variances not assumed			.699	93.895	.487	1.483	2.122	-2.732	5.697
FallPSS	Equal variances assumed	.455	.501	1.669	94	.098	4.203	2.518	-.796	9.203
	Equal variances not assumed			1.664	91.662	.100	4.203	2.526	-.814	9.221
SpringPSS	Equal variances assumed	2.985	.087	2.575	94	.012	7.712	2.995	1.765	13.659
	Equal variances not assumed			2.561	89.926	.012	7.712	3.012	1.729	13.695

4.The sample problems were investigated by students in the flipped class group.

a. Crude salt is salt produced from seawater or boiled in salt wells. Once the insoluble impurities, such as sand and silt, have been removed, and only then is the salt referred to as refined salt. How to convert coarse salt to refined salt in an experiment?

b. Is the mass of the candle before burning equal to the mass after burning plus the mass of the combustion products in a confined environment?