Appendix 1: Details of PSL interventions.

Three Institutions participated in this study. Two institutions followed a PSL interventions with their students. The control group read no PSL as part of their course. Both treatment groups began their semester with an introductory class session on how to read PSL, how PSL and science literacy are related, and the differences between a “good” and “bad” PSL. Both instructors at Institution 1 and 2 used identical teaching strategies for science literacy lessons. Instructors at PSL1 and PSL2 assigned three PSL’s at three intervals during a 15-week semester. Papers correlated to content taught in class by the instructor. Students were assigned to break out rooms in zoom along with guided questions for paper discussion. The breakout room sessions were monitored by teaching assistants and the instructor. This intervention was developed by the research team and has not been previously tested.

Instructors selected PSL based on content being taught in their course. Additional suggestions for selecting appropriate PSL can be found in Chatzikyriakidou et al., 2021. The assigned pieces of PSL are as follows (colored boxes represent the same PSL being used in both interventions). The PSL selected for this study is short (~3 pages) and directly connects to biology content being taught in the course, making this an appropriate intervention for novice PSL readers.
PSL 1 students completed PSL readings and assignments entirely at home. Students were given a PDF of the PSL paper and guided reading questions as a graded assignment. Students were allowed to post responses and engage in a discussion thread using Blackboard. Students were given a formative assessment science literacy quiz related to the assigned PSL.

PSL 2 students completed PSL readings at home with discussion during class (virtual). Similar to PSL 1 students read the PDF version of the PSL using identical guide questions as PSL 1. However, questions were not graded and used as a summative assessment. Students discussed answers to the guided reading questions in breakout groups and submitted answers using a student response system for participation points as part of their course grade. Following breakout discussions, the instructor reviewed students answers and took a short quiz at the end of class related to the assigned PSL.
Appendix 2: Additional methods for descriptive statistics.

The normality of the data was found to be within the acceptable range of skewness and kurtosis (-2 and 2) (George & Mallery, 2019). Mahalanobis distance was also measured as a means of detecting any potential outliers. Potential outliers would have a high Mahalanobis distance with a level of significance for this was p<0.001. After screening the data and removing outliers the resulting data set contained 322 data points from the pre-test. The factorability of the 322 data set was then tested using the Keiser-Meyer-Olkin (KMO) measure for sampling adequacy. The resulting score was a 0.817, which is above the previously established threshold, and which demonstrates that the sample is large enough to be reliable, and will suit the factor analysis (Kaiser, 1975). Following this, Bartlett’s Test of Sphericity was conducted, and determined that the correlations, collectively, met the level of significance at 0.001. Thus, the established methodology is appropriate for the data. Finally, all anti-image correlations, as determined by the KMO measures for each individual variable, fell within the acceptable range, by exceeding the threshold of 0.50.
Appendix 3: EFA additional methods

The goal of EFA is to determine whether questionnaire items have adequately measured the intended variable and are useful in measuring tools for the study’s hypotheses and related research questions. Thus, it is considered a best practice for validation of measurement (Watkins, 2018). As an extraction method in EFA, Oblique (Geomin) Rotation with maximum-likelihood with robust standard errors (MLR) (UCLA, 2022), which is useful in studying the correlations between factors in the given output, was chosen.

Visual inspection of the scree plot (Appendix 3), parallel analysis based on eigenvalues from the principal components, and factor analysis in combination with theoretical considerations were used to decide on the appropriate number of factors to retain (psych package; Revelle, 2017).

To verify the factorability of the data, a correlation matrix was developed using Bartlett’s test of sphericity (Bartlett, 1954). Bartlett’s test for sphericity is considered statistically significant if the resulting $p < 0.05$ (Hadi et al., 2016). Specifically, this ensures that the correlation matrix of the variables, or measurements from each item of the questionnaire, are significantly divergent when compared to the identity matrix (Bartlett, 1954; Hadi et al., 2016). If the difference is found to be statistically significant, then data reduction techniques can be meaningfully used on the data, or the data is considered valid (Pallant, 2020; Hadi et al., 2016).

Then, sampling adequacy was tested according to the Kaiser-Meyer-Olkin (KMO) measure (Kaiser, 1974). KMO considers the sample adequate only if the KMO is greater than .5 (Kaiser 1974; Field 2000). Generally, results between .5 and .7 are considered mediocre, values between .7 and .8 are considered acceptable, and values over .8 are considered superior or preferred. Thus, within the current study, we will accept sampling adequacy only of KMO is greater than .7 (Pallant, 2020; Kaiser 1974).

If the extracted factors and the factorability of the data were confirmed, there are criteria to be met for making decision for the number of factors to retain. The traditional cut-off value of eigenvalue is 1.0 (Tabachnick & Fidell, 2007). And the rotated factor loadings that are not below 0.4 were retained and for the cross-loaded factors, the highest loading to interpret factor was selected, when the absolute difference was > 0.2 (Stamper & Masterson, 2002). This procedure was repeated until the all the criteria was successfully met.
Scree Plot of the EFA model for the Thinking Tools Assessment
Appendix 4: CFA methods

CFA is a structural equation modeling technique used to determine the goodness of fit between a hypothesized model and the sample data. To determine the model fit of CFA, there is the criteria of indices to see the model is fitted to the data, including chi-square, root-mean-squared error of approximation (RMSEA), comparative fit index (CFI), Tucker-Lewis Index (TLI) and Standardized Root Mean Square Residual (SRMR). When comparing models, a lower chi-square value indicates a better fit, given an equal number of degrees of freedom, however, chi-square is sensitive to sample size. RMSEA is a measure of the average of the residual variance and covariance (RMSEA: < .05 good, <0.08, 0.1 marginal, >.1 poor; Brown & Cudeck, 1993). CFI is an index that fall between 0 and 1, with values greater than 0.90 considered to be indicators of good fitting models (Bentler, 1990). Tucker Lewis Index (TLI; >.90 acceptable, >.95 excellent; Tucker & Lewis, 1973. The SRMR is an absolute measure of fit and is defined as the standardized difference between the observed correlation and the predicted correlation. A value less than .08 is generally considered a good fit (Hu & Bentler, 1999). CFA considers each element of the measurement tool that has been developed, and assesses it for dimensionality, reliability and validity (Said et al., 2011). Dimensionality is determined by factor loading and is confirmed or accepted if all items have a factor loading greater than 0.5 (Hair et al., 2010).
Appendix 5: Additional EFA results

The result of the initial analysis of EFA on the items in Table 1 revealed a probable three-factor solution with Oblique (Geomin) Rotation using maximum-likelihood with robust standard errors (MLR) (UCLA, 2022). This is useful in studying the correlations between factors in the given output. The eigenvalue for the first factor was 4.533, for the second factor was 1.499, and for the third factor was 1.191. No other factors had eigenvalues > 1. Based on the structure established by Field (2013), the research suppressed all factor loadings which did not meet the threshold, established at 0.3. Those scores failing to meet the threshold were removed. Scores 0.4 and greater were, therefore, considered stable (Guadagnoli and Velicer, 1988). Cross loading occurs in a case where the variable is found to have multiple significant loading factors. If a label or factor is shared in multiple factors, it can be more difficult to separate the factors into clear, distinct concepts (Hair et al., 2010; Le & Cheong, 2010). Then, the highest loading to interpret factor was selected, when the absolute difference was > 0.2 (Stamper & Masterson, 2002). This protected the factor interpretation from cross-loading.

Rotated factor loading of the EFA model for the PSL Reading Strategies

| item | Factor 1: | Factor 2: | Factor 3: |
|------|-----------|-----------|-----------|
| 2    | .687      |           |           |
| 3    | .855      |           |           |
| 4    |           | .851      |           |
| 5    |           | .838      |           |
| 6    |           | .592      |           |
| 9    |           |           | .567      |
| 10   |           |           | .995      |
| 11   |           |           | .714      |
References:

Bartlett, M. S. (1954). Significance test for sphericity of a normal n-variate distribution. *Journal of the Royal Statistical Society*, 16, 296-298. Hadi et al., 2016

Bentler, P. M. (1990). Comparative Fit Indexes in Structural Models. *Psychological Bulletin*, 107, 238-246. http://dx.doi.org/10.1037/0033-2909.107.2.238

Brown, M. W., & Cudeck, R. (1993). Alternative Ways of Assessing Model Fit. In K. A. Bollen, & J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 136-162). Newbury Park, CA: Sage.

Byrne, B. M. (2013). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. Routledge.

Field, A. (2000). Discovering statistics using SPSS for Windows: advanced techniques for beginners. *Great Britain: Sage Publication*.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Sage.

George, D., & Mallery, P. (2019). *IBM SPSS statistics 26 step by step: A simple guide and reference*. Routledge.

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological bulletin, 103*(2), 265.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational research methods, 7*(2), 191-205.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

Kaiser, H. F. (1974). An index of factorial simplicity. *psychometrika, 39*(1), 31-36.

Kaiser, H. F., & Michael, W. B. (1975). Domain validity and generalizability. *Educational and Psychological Measurement, 35*(1), 31-35.

Klinke, S., Mihoci, A., & Härdle, W. (2010). Exploratory factor analysis in MPlus, R and SPSS. *Invited paper ICOTS8 of the International Association of Statistical Education*.

Le, T. C., & Cheong, F. (2010). Perceptions of risk and risk management in Vietnamese Catfish farming: An empirical study. *Aquaculture Economics & Management, 14*(4), 282-314. https://doi.org/10.1080/13657305.2010.526019

Pallant, J. (2020). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS*. Routledge.
Sommet, N., & Morselli, D. (2021). Keep calm and learn multilevel linear modeling: A three-step procedure using SPSS, Stata, R, and MPlus. *International Review of Social Psychology, 34*(1).

Stamper, C. L., & Masterson, S. S. (2002). Insider or outsider? How employee perceptions of insider status affect their work behavior. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior, 23*(8), 875-894.