# Assessing the Science Learning Assessment (SLA) Instrument in Greek Early Childhood Education Using the Item Response Theory Framework

Vasilis Grammatikopoulos[1]*, Nikolaos Tsigilis[2] and Athanasios Gregoriadis[2]

[1] International Hellenic University, Thessaloniki, Greece, [2] Aristotle University of Thessaloniki, Thessaloniki, Greece

Studies revealed that science learning in early childhood is associated with primary and secondary school readiness, improved causal reasoning, increased interest in science in later life, and with long-term effects on academic achievement. Thus, having valid and reliable instruments to assess children's science learning becomes critical in early childhood education (ECE). The main goal of the current study was to examine the psychometric properties of the Greek version of the Science Learning Assessment (SLA), an instrument developed in the USA to assess science learning in ECE. Randomly selected from 53 preschool units in Greece, 528 children were administered with the Greek version of the SLA. Advanced statistical methods (CFA, IRT framework) were employed to analyse the data and the results showed that the Greek version of the SLA displayed acceptable psychometric properties. Therefore, it can be argued that the Greek version of the SLA is a valid and reliable scale for assessing children's science knowledge in ECE, capturing universal concepts of science knowledge children in early childhood.

Keywords: early childhood education, educational assessment, item response theory, psychometric properties, science learning assessment

## INTRODUCTION

Science learning is a very important subject in education, and it is anticipated that future demand for science knowledge will increase due to emerging challenges in the fields of feeding, growing population, spread of diseases, water supply, climate change etc. The demand for more experts in several science subjects, and for a higher level of citizens' scientific literacy is constantly increasing, since there is a growing need to improve the understanding of all these socio-scientific topics that have a substantial impact in human lives (Hammer and He, 2014).

### Science Learning in Early Childhood

In contrast with the Piagetian perspective that children in early childhood have limited skills of learning science due to developmental limitations, it is now widely acknowledged that science learning holds a critical position in early childhood education (ECE) as children are very competent in capturing complex ideas from early years (Metz, 2004; Akerson et al., 2015; Clements and Sarama, 2016). Results from recent studies in Greece (Kallery et al., 2009; Ergazaki et al., 2015; Kalogiannakis et al., 2018; Kalogiannakis and Papadakis, 2019), and other countries examining science knowledge in early childhood, revealed that children's eagerness and curiosity in learning

constitute an ideal field where science knowledge can be cultivated (Gelman and Brenneman, 2004; Brenneman and Louro, 2008; Samarapungavan et al., 2008, 2009, 2011; Mantzicopoulos et al., 2009; Bagiati et al., 2010; French and Woodring, 2012; Ergazaki et al., 2015; Greenfield, 2015; Guo et al., 2015; Vitiello et al., 2019). Thus, concerns about the developmental readiness of preschool children for science knowledge can be overridden, since children's ability to internalize knowledge about science concepts has been proven. Therefore, it can be argued that it would be most beneficial to deal with science content knowledge from this early stage of their education.

In line with the previous arguments, several studies showed that science learning in ECE is associated with primary and secondary school readiness, improved causal reasoning, increased interest in science in later life, and long-term effects on academic achievement (Clements, 2001; Watters et al., 2001; Arnold et al., 2002; Connor et al., 2004; Ginsburg and Golbeck, 2004; Kallery, 2004; Hamre and Pianta, 2005; Duncan et al., 2007). Brain and neuroscience research revealed that, for certain domains, learning occurs most efficiently in a time period early in children's life, and science learning is one of these domains (Kallery et al., 2009). Children at this age are motivated to learn science content, and the lack of encouraging them toward this direction is likely to result in a decline of their curiosity and interest (French, 2004; Eshach and Fried, 2005). Hammer and He (2014) also supported the recent research-based evidence, by suggesting that ECE is an appropriate environment for emerging children's awareness in science.

Introducing science teaching in ECE is not considered that will -ad hoc- enhance children's knowledge and interest. Content, didactical approaches, and teacher competencies are very crucial for the successful implementation of science teaching in ECE. Recent studies revealed that many problems in promoting science learning in ECE can be attributed to teachers' attitudes and/or competencies (Gomes and Fleer, 2018; Sundberg et al., 2018).

## Science Learning in the Greek ECE Setting

As in many ECE systems around the globe, in the Greek ECE science teaching is constantly gaining attention the last two decades. First, the national curriculum for the Greek ECE developed in 2002 and implemented since then in all ECE settings discusses in detail the field of sciences, including activities in the domains of physics, biology, environment etc. (MoE/PI, 2002). The Greek national curriculum acknowledges the crucial role of children's engagement with science activities and supports their physical and mental involvement with science activities (Kallery et al., 2009).

All of the academic departments responsible for ECE have already developed relevant modules at under- and postgraduate level for their students. These academic departments have also organized several conferences the last 20 years (see for more details in sece.gr), covering extensively the field of science in ECE (e.g., curricula, attitudes, learning environment, didactical approaches for science teaching, environmental issues, etc.). This scientific activity has boosted research in the Greek ECE settings regarding the science topic, resulting to several publications in international peer reviewed journals (e.g., Kallery et al., 2009; Giallousi et al., 2014; Fragkiadaki et al., 2017; Kalogiannakis et al., 2018; Kanaki and Kalogiannakis, 2018; Kalogiannakis and Papadakis, 2019). Conclusively, it can be argued that the domain in Greece has gained at least the attention of the researchers and academics.

## Assessing Science Learning in Early Childhood

As science education in ECE is expanding, new curricula, or programs are developed and implemented. This expansion raises the issue of assessing accurately children's gains in knowledge from such initiatives. Researchers pointed out that the increased focus on science education in ECE is not accompanied by a similar focus on the assessment field, revealing the dearth of research in validly evaluating science knowledge in ECE (Greenfield, 2015; Vitiello et al., 2019). A possible explanation could relate to the fact that unlike other areas such as literacy, the creation of universal standards is lacking for science education in ECE (Greenfield, 2015). To ensure the merit and worth of science curricula or programs, measures with sound theoretical and psychometric characteristics are essential. However, researchers stressed the lack of an adequate number of reliable and valid instruments available for assessing early childhood science knowledge (Brenneman, 2011; Kloos et al., 2012; Greenfield, 2015; Zucker et al., 2016). In the Greek ECE this lack is considered more extended, as despite the increase of relevant studies in the Greek research community, evidence for the psychometric properties of instruments are not in the core of the research questions addressed (e.g., Tsigilis et al., 2007; Grammatikopoulos et al., 2008, 2015; Grammatikopoulos, 2012).

The instrument used in the current study is called Science Learning Assessment (SLA), and it has been developed to evaluate early childhood student's knowledge in science (Samarapungavan et al., 2008, 2009, 2011; Mantzicopoulos et al., 2013). It consists of two dimensions, the Scientific Inquiry Processes (SIP) factor containing nine items, and the Life Science Concepts (LSC) factor containing 15 items.

The establishment of the validity and reliability of an instrument is a long and ongoing procedure (Zhao and Gallant, 2012). The present study is an attempt to test the function of an instrument measuring science knowledge in ECE in a different country than the one where it was initially developed. An additional challenge to face, is that there are findings not fully confirming the factor structure of the SLA (Samarapungavan et al., 2009) and the authors suggested that the instrument has to be further tested in the field.

To the best of our knowledge the appropriateness of the SLA has been examined solely within the Classical Test Theory (CTT), namely through exploratory factor analysis and internal consistency. Despite the fact that CTT is a valuable approach, there is criticism for its limitations, such as sample dependent statistics, inability to discriminate between item difficulty, and participants proficiency, identical measurement error for all scores etc. Item response theory (IRT) is a modern measurement framework for examining the psychometric properties of tests

and questionnaires. IRT has many advantages in comparison to CTT (Embretson and Reise, 2000; DeMars, 2010). Contrary to CTT which focuses on a total test score, IRT models relate items performance to the underlying trait using a non-linear function. In addition, IRT can discriminate between item difficulty and participants' ability levels. When its assumptions are satisfied, item parameter estimates are independent of the specific sample which has been used to calculate them. The inherent measurement advantages of the IRT over CTT make it a valuable alternative, which reveals aspects of items, measures, and concepts that CTT cannot unravel (Singh, 2004).

## Purpose of the Study

The purpose of the current study was to examine the initial psychometric properties of the Greek version of the SLA using the Item Response Theory framework and to test its invariance across gender. The main research question of the study focused on whether the Greek version of the SLA has adequate psychometric properties for assessing science knowledge in the Greek ECE.

## METHOD

### Participants

The random selection of ten children per classroom from 53 Greek public preschool classrooms resulted to 528 children (271 boys, 257 girls) who participated in the current study. Their mean age was 60.3 months (± 16.1). The parent of each participating child was administered and completed a written consent form.

### Measures and Procedure

The instrument of the current study was the SLA, a measure of science learning for ECE students (Samarapungavan et al., 2008, 2009, 2011; Mantzicopoulos et al., 2013). The SLA consists of 24 items designed to assess target concepts of science knowledge based on the National and State (Indiana) Science Education Content Standards for Kindergarten. Nine items measure the SIP (e.g., understand the empirical basis of science, knowledge about the natural world), and 15 items measure the LSC (e.g., understand the characteristics of living things, understand living things' life cycles) (Samarapungavan et al., 2009). The SLA response process requires from each respondent to select among multiple choices the right one. Respondents are administered two, three or four choices for every question, and they have to choose the right one (**Figure 1**). Thus, the SLA scoring method is based on a binary format in which correct answers are given a score of 1 and wrong answers or non-answers a score of 0, and the possible total score on the SLA ranges between 0 and 24.

In previous research, the SLA was tested for its psychometric properties using Confirmatory Factor Analysis (CFA) showing a non-tenable structure (Samarapungavan et al., 2009). For the Greek version of SLA, the authors decided to include all the items of the original scale, but responses to three pairs of them (10–11, 12–13, 20–21) were combined, as the items 10, 12, and 20 were just "basis" questions that could be correctly answered by chance, whereas correct responses for the followed items (11, 13, and 21, respectively), again in correct/wrong format, were justifications for the preceded items (see **Figure 1**). The original SLA was back translated and then it was presented to an expert (experienced academic scholar in science education) in order to secure that the items of the Greek version of the SLA were meaningful in Greek language and easily understandable by children in early childhood.

Following the procedure described in the study of Samarapungavan et al. (2009), the children were administered the scale individually in the teacher's office and the whole procedure for the completion of the SLA lasted ~10 min.

## Data Analysis Strategy

To examine the dimensionality of the SLA confirmatory factor analysis (CFA) procedures were employed. Following the work



One of these aminals is an insect (show the pictures)
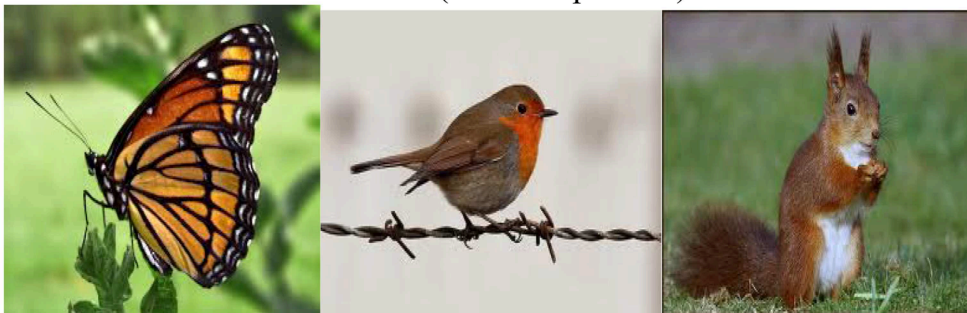
10. Which one? (a) Butterfly; (b) Bird; (c) Squirrel (Correct answer: Point or say 'the butterfly')
11. (follow up to 10 in case of correct answer) How do you know that this is an insect? (Correct answer: Six legs)

**FIGURE 1 |** Schematic representation of the items 10 & 11 of the Greek version of SLA.

of Samarapungavan et al. (2009), two candidate models were postulated. The first model hypothesized that one latent factor underlies responses to 21 items, whereas the second model hypothesized a two-factor structure.

All analyses were conducted using *Mplus* ver. 7.3 (Muthen and Muthen, 2012). Given that SLA's responses are dichotomous, variables declared as categorical in the *Mplus* syntax meaning that a tetrachoric correlation matrix was calculated and entered for analysis. Moreover, the mean and variance adjusted Weighted Least Squares (WLSMV) was used as the most appropriate estimator for the specific data set (Wang and Wang, 2012). Evaluation of model's fit was based on $\chi^2$ statistic, Comparative Fit Index (CFI), and Root Mean Square Error of Approximation (RMSEA). CFI values around 0.95 and RMSEA values of 0.06 were considered as indicative of a well-fitting model (Hu and Bentler, 1999). With categorical data *Mplus* provides the weighted root mean square residual (WMSR) instead of the standardized root mean square residual (SRMR). WRMR values of 1.00 or lower denote a good fit (Yu, 2002).

Based on the CFA results the most viable model, unidimensional or multidimensional, was selected to calibrate SLA items. Likelihood ratio test results determined the appropriate number of parameters. Next, it was examined whether SLA latent scores are independent of children gender. In case that boys and girls with the same latent score have different probabilities for selecting the correct response for a particular item, then this item exhibits differential functioning (DIF). Additionally, it has to be pointed out that, when translating a questionnaire into another language, the psychometric properties of the items may change. Thus, researchers suggest examining the equivalence of the items using Differential Item Functioning (Penfield and Camilli, 2007; Greenfield, 2015). Differential item functioning (DIF) was tested within the structural equation modeling framework using multi-group approach (Finch and French, 2007; Muthen and Muthen, 2012). Initially, the most tenable SLA model was fit separately to boys and girls. Following this step three consecutive models were tested, in which constraints were introduced in a hierarchically increasing fashion. Configural model (simultaneously fit to both gender with no additional constraints, M0), metric invariance model (equal loadings, M1), and scalar invariance model (equal thresholds, holding loadings invariant, M2). Presence of metric invariance suggests equal item discrimination and hence non-uniform DIF is not an issue. On the other hand, if scalar invariance holds, it means that both discrimination and location parameters are equal, that is absence of DIF. Because the examined models were nested $\Delta\chi^2$ was used to compare the unconstrained to the (more) constrained model. Given that the estimator was the WLSMV, the DIFFTEST option provided by Mplus was utilized for model comparisons.

## RESULTS

### Descriptive Analysis

Participants' descriptive data regarding their SLA scores, for the overall sample as well-across gender are presented in **Table 1**. It is interesting to notice the relatively high score the Greek students

**TABLE 1** | Descriptive statistics for the Science Learning Assessment of the Greek sample.

| SLA | Overall | Boys | Girls |
|---|---|---|---|
| SLA—Scientific Inquiry Processes | 5.27 (1.99) | 5.28 (2.05) | 5.85 (3.94) |
| SLA—Life Science Concepts | 5.94 (3.81) | 5.58 (1.98) | 6.02 (3.70) |
| SLA—Total | 11.37 (5.34) | 11.13 (5.45) | 11.60 (5.24) |

*Means represent average number of correct answers of the SLA items.*

in the SLA—Scientific Inquiry Processes subscale (5–6 correct answers out of 9) achieved, and the low score in the SLA—Life Science Concepts subscale (5–6 correct answers out of 15).

### Dimensionality of the SLA

CFA results concerning the unidimensional model showed an issue attributed to item 23 (items retain the same numbering of the original SLA). Specifically, the residual covariance matrix was not positive and item's 23 standardized factor loading exceeded unity. This item was excluded, and the analysis was rerun. Although $\chi^2$ value was statistically significant, goodness-of-fit indices suggesting an excellent fit to the data, $\chi^2 = 292.96$, $df = 170$, $CFI = 0.973$, $RMSEA = 0.037$, 90%CI $RMSEA = 0.030$ to 0.044 $WRMR = 1.080$. Similar fit indices were observed for the two-factor model, $\chi^2 = 268.70$, $df = 169$, $CFI = 0.978$, $RMSEA = 0.033$, 90%CI $RMSEA = 0.026$ to 0.041, $WRMR = 1.022$. Chi-square difference test between the two models showed that the two-factor model yielded a better fit to the data than the unidimensional model ($\Delta\chi^2 = 13.89$, $df = 1$, $p = 0.0002$). All factor loadings were statistically significant ranging from 0.31 to 0.90. The association between the two latent factors was positive and very strong (0.874, $SE = 0.033$). A 95% confidence interval around the estimated correlation did not include unity, suggesting that although "Scientific Inquiry Processes" (SIP) and "Life Science Concepts" (LSC) factors are highly correlated they represent distinct constructs. Thus, the two-factor model was selected as the most tenable for the SLA data.

### Item Response Analysis of the SLA

Next, a multidimensional IRT model was fitted to the SLA items. A simple structure model was adopted, in which an item was related only to its corresponding dimension (Hartig and Höhler, 2008). In order to select the most appropriate multidimensional IRT model, we fitted 1PL ($-2LL = 10664,72$; $df = 23$; AIC = 10706,72; BIC = 10796,45), 2PL ($-2LL = 10255,99$; $df = 41$; AIC = 10337,99; BIC = 10513,18), and 3PL ($-2LL = 10340,46$; $df = 61$; AIC = 10462,46; BIC = 10723,10) models. All models were converged without any problems. The $-2LL$ as well as AIC values suggested that the best fitting model was the 2PL. Moreover, likelihood ratio tests also indicated the prevalence of the 2PL model. Examination of items local independence for the 2PL model showed that standardized LD $\chi^2$ values were below 10, suggesting that this assumption was met (Cai et al., 2011). The largest association after accounting for the multidimensional IRT model was between items #Life 19 and #Life 21 (standardized LD $\chi^2$ values = 8.4). To test SLA fit at item level the S-$X^2$ (Orlando

**TABLE 2 |** SLA items calibration results.

|  | Discrimination | Threshold |
|---|---|---|
| SIP 1 | 1.04 | −1.80 |
| SIP 2 | 0.65 | 0.00 |
| SIP 3 | 0.53 | 0.01 |
| SIP 4 | 1.29 | −0.87 |
| SIP 5 | 0.60 | 0.46 |
| SIP 6 | 1.11 | 0.85 |
| SIP 7 | 1.32 | 0.07 |
| SIP 8 | 1.75 | −0.54 |
| SIP 9 | 2.51 | −1.38 |
| LSC 11* | 1.47 | 3.85 |
| LSC 13* | 1.47 | 3.76 |
| LSC 14 | 2.14 | 0.39 |
| LSC 15 | 1.67 | 2.82 |
| LSC 16 | 2.99 | 0.70 |
| LSC 17 | 2.26 | 1.41 |
| LSC 18 | 2.40 | 1.73 |
| LSC 19 | 2.16 | 1.55 |
| LSC 21* | 2.46 | 2.21 |
| LSC 22 | 4.01 | 1.27 |
| LSC 24 | 1.90 | 1.91 |

*Combined items.

and Thissen, 2003) was employed. *P*-values were adjusted using Holm-Bonferroni method. Results showed that two items (#Life 22 and #Life 24) yielded statistically significant values.

**Table 2** presents the derived multidimensional IRT parameters. The location estimates for the Scientific Inquiry Processes dimension ranged from 0.46 to −1.80 suggesting that the items were fairly easy for the participants. Moreover, the discrimination estimates ranged from 0.53 to 2.51 indicating that there was a substantial variability in the degree to which the items discriminate. With regard to the Life Science Concepts dimension location estimates varied from 0.39 to 3.85 suggesting that most of the items were rather difficult for the children. In addition, discrimination estimates yielded high values indicating that these items could effectively discriminate among participants with similar knowledge in Life.

Test information of the SIP and LSC are presented in **Figures 2**, **3**, respectively. Information is not constant but fluctuates in relation to the ability level. Test discrimination ability is maximized in areas of the latent trait that has the highest information level (area around the pick of the information curve). SIP information yielded its highest value about half standard deviation below average. From **Figure 2** it can be easily inferred that this SLA dimension reliably captures children with SIP within a −1.5 to 0.05 range. Information concerning the LSC dimension peaks about half standard deviation above average and it seems to reliably assess children within a −0.5 to 1.5 range. Although a single number for reliability it is not frequently reported in IRT for item or test information, we decided to present it in order the current study's results to be comparable

with previous studies. Using the *mirt* library in R the values were: 0.804 for SIP dimension and 0.841 for LSC dimension.

After item calibration, children responses to SLA items were examined for metric and scalar invariance. DIF results are presented in **Table 3**. First, the derived two-factor model was simultaneously fit to both groups (M0). Chi-square values were statistically significant for males but not for females. However, all goodness-of-fit indices suggest an excellent fit to the data. Next, the Configural model was tested, in which the 2PL model was simultaneously fit to both genders with no constraints. Results clearly denote the tenability of similar SLA construct across gender. Thus, the necessary condition for testing differentiation of item parameters was satisfied.

At the following step of analysis item loadings were constrained to be equal across gender. Chi-square value was not statistically significant and goodness-of-fit indices satisfied the cut-off values. Moreover, $\Delta \chi^2$ was not statistically significant, suggesting that the addition of equality constraints did not deteriorate the fit of the model. Thus, the hypothesis of equal items loading was met, indicating that the SLA item discrimination parameters are stable across gender.

Finally, the scalar invariance model was considered (M2), in which the equality of items threshold was tested, holding items loading invariant. Despite the excellent fit of the model, $\Delta \chi^2$ was statistically significant. Modification indices were used to locate the source of model's deterioration. Results showed that the first item of the SIP dimension might function differently for boys and girls. Thus, threshold of item 1 was allowed to vary across gender and the model was rerun (M2mod). A comparison between the metric invariance model with the modified scalar invariance model yielded a non-significant model, suggesting that, with exception of item 1, SLA items difficulties are robust across gender. Examination of the thresholds for item 1 showed that boys yielded higher values (−0.741) in comparison to girls (−1.434).

## DISCUSSION

A notable finding of this study was the performance the Greek children scored on the SLA. Their total mean score ranked at 11.37 (out of 24). Comparing their performance with the children from USA (12.03/24) that participated in the Samarapungavan et al. study (2009), it was found that both groups were relatively at the same level of knowledge ($p = 0.47$). However, a separate examination and interpretation of the scores on the two subscales of the SLA can lead to quite different conclusions. The Greek group scored on the Scientific Inquiry Processes quite high (5.27/9), while the USA control group scored 2.89 ($p < 0.001$) (Samarapungavan et al., 2009). The situation was quite different regarding the Life Science Concepts. The Greek group scored 5.94/15) whereas the USA control group scored 9.14 ($p < 0.001$).

Samarapungavan et al. (2009) argued that children are likely to acquire knowledge about nature without understanding the processes, while the converse is unlikely. In the current study, the findings revealed the opposite. Greek children seem to know less
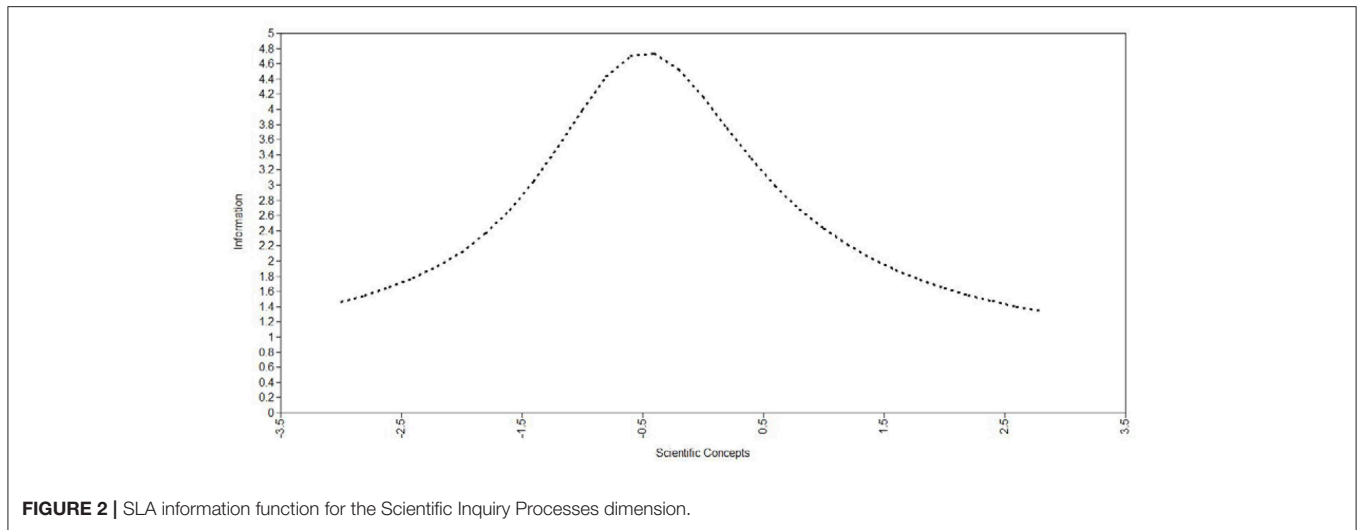
**FIGURE 2 |** SLA information function for the Scientific Inquiry Processes dimension.
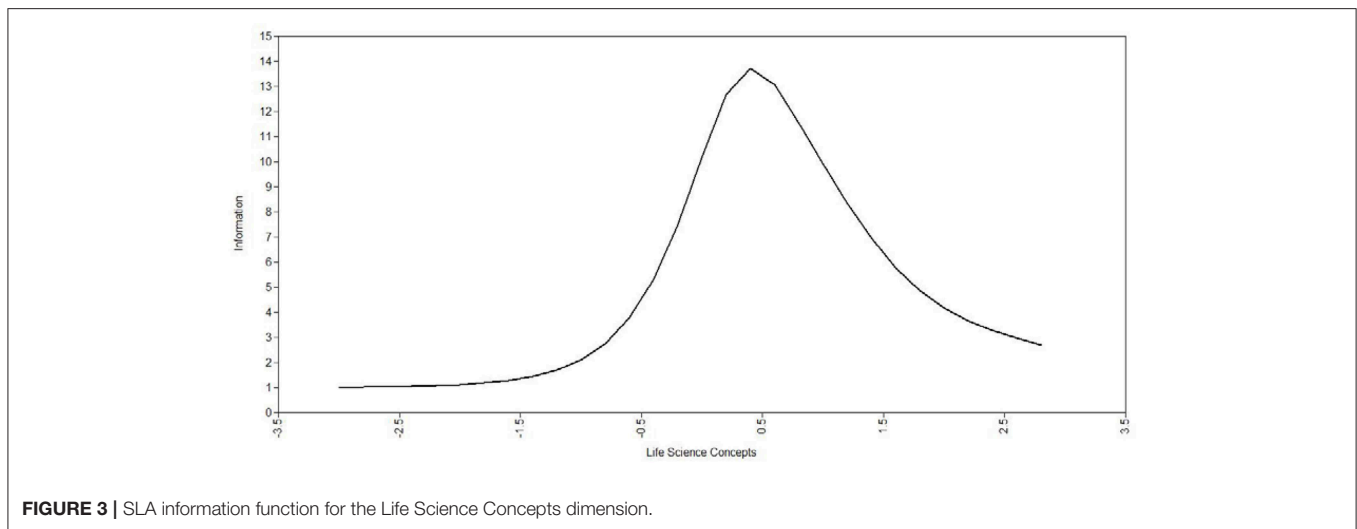


**FIGURE 3 |** SLA information function for the Life Science Concepts dimension.

about nature than about the processes by which this knowledge is constructed.

The reason for this finding could be partially attributed to the curriculum of the Greek ECE. The Greek national curriculum focuses mainly on processes like the content of the Scientific Inquiry Processes items, and not on nature or environment knowledge that are the content of the Life Science Concepts items. An additional explanation could be found at the different environment of the studies. The most characteristic evidence was the scores on the item 6 [(Show pictures) Two girls found an egg. "the girl in green thinks it is a duck egg. "the girl in blue thinks it is a goose egg. How can they find out what it is?]. This item was the most difficult to be answered correctly (2 and 1.7% correct answers) in two studies in the USA (Samarapungavan et al., 2009, 2011), while it appeared to have moderate levels of difficulty for the Greek participants (see **Table 2**). Another possible reason for the above result could be attributed to the aggregation of the three pairs of items (all of them were in the LSC factor). In the Greek version of the SLA these items were by far more difficult than the

**TABLE 3 |** Chi-square and goodness-of-fit indices for the DIF analysis.

|  | $\chi^2$ | $\Delta\chi^2$(df) | CFI | RMSEA | 90%CI RMSEA | WRMR |
|---|---|---|---|---|---|---|
| M | 208.5 (169)* | – | 0.984 | 0.029 | 0.012–0.042 | 0.876 |
| F | 196.5 (169)$^{ns}$ | – | 0.985 | 0.025 | 0.001–0.039 | 0.862 |
| M0 | 404.7 (338)* | – | 0.985 | 0.027 | 0.015–0.037 | 1.23 |
| M1 | 389.0 (356)$^{ns}$ | 7.2 (18)$^{ns}$ | 0.992 | 0.019 | 0.001–0.030 | 1.26 |
| M2 | 418.6 (374)$^{ns}$ | 37.8 (18)* | 0.990 | 0.021 | 0.001–0.021 | 1.31 |
| M2mod | 405.6 (373) | 16.4 (17)$^{ns}$ | 0.993 | 0.018 | 0.001–0.029 | 1.28 |

*M, Males; F, Females; M0, Configural model; M1, Metric invariant model; M2, Scalar invariant model.*
*$^*p < 0.05$.*
*ns, no significant.*

USA version, as the Greek students had to answer both questions of each pair correctly. Of course, further directly comparative studies in different environments would be required in order to draw solid conclusions for this case.

The results in our study confirmed the two-factor structure of the SLA as initially was designed (Samarapungavan et al., 2009), supporting the research question. The 20-item Greek version of the SLA appeared with two robust factors (SIP & LSC) with nine and 11 items, respectively. The results of this study about the Greek version of the SLA argued that it can be considered a valid and reliable instrument for evaluating science knowledge in the Greek ECE. The validity of the Greek version of the SLA can be inferred by the factor structure and the adequate discrimination parameters (**Table 2**). The precision of measurement however seems to differentiate in relation to specific factors. In particular, SIP appears to assess with reasonable levels of precision children with low to average levels of science knowledge (**Figure 2**). On the other hand, LSC assesses precisely, children with average to high levels of science knowledge (**Figure 3**).

The IRT framework, employed in the current study provided interesting evidence for the function of the specific items. First, difficulty parameters indicated that the SIP was not so difficult for Greek children as it was for the USA sample in Samarapungavan et al. studies (2009; 2011). The Greek participants seem to understand conceptually the key aspects of scientific processes as measured by the Greek version of the SLA. Regarding the LSC items IRT analysis revealed that the LSC items were quite difficult to be answered. Samarapungavan et al. (2011) argued that such findings are vital in the light of developmental research investigating children's limitations in their understanding of the science learning in early school years. Second, discrimination parameters indicated that the SLA items effectively discriminate Greek participants in our study. Moreover, it has to be pointed out that Life Science Concepts items yielded higher disclination parameters in relation to the Scientific Inquiry Processes items (**Table 2**). These results were not in congruence with the analyses investigating discrimination parameters in the Samarapungavan et al. study (2009). Yet, a direct comparison cannot be applied, as the method used then (Kelley's procedure, 1939) differs from the more sophisticated and advanced IRT framework that was applied in the current study.

For the factor Scientific Inquiry Processes all the items (9) were retained, whereas one item (#23) was discarded from the Life Science Concepts dimension based on the analysis. There isn't any apparent reason why this item was proven problematic in the current study. In Samarapungavan et al. (2009) study this item appeared as the easiest item of the instrument (95% correct answers). Moreover, when translating items into another language it is likely that some of them might change their properties (Penfield and Camilli, 2007; Greenfield, 2015).

Three pairs of items (#10–11, #12–13, #20–21) of the questionnaire were aggregated for conceptual reasons. The items 10, 12, and 20 are basis questions for the followed items 11, 12, and 21. For example, *item 20: "Which of these is a living thing (show pictures)? a) Plant; b) Car; c) Table"* Then, after the participant's answer the question 21 follows: *"Why is it a living thing? How can we tell that it is a living thing?"* Here, the correct response is considered if the participant names two or more characteristics of living things, e.g., grows, needs food/water, breathes, moves on its own, etc. As it is obvious, the key response here for the understanding of the Life Since Concept is the

justification of why *a Plant* is a living thing, and not just a guess of which a living thing is. Thus, responses were combined for the above three pairs. Only if the participant correctly answered the first question of each pair, we continued with the next one, and if they also correctly answered the second part, only then a positive score was awarded for the item. As it was expected these items yielded increased difficulty levels in comparison to rest of the LSC items. Moreover, the results justified the above decision as the proposed two-factor structure was replicated in the current study.

The LSC factor needs to be further studied, especially in the Greek ECE, where the focus of the teaching is closer to the content of the Scientific Inquiry Processes items. It would also be interesting for future studies if an intervention was designed and implemented focusing on the development of children knowledge about nature, environment and relevant concepts.

With the exception of the first question, no differences were detected between boys and girls regarding their knowledge in science as assessed by the SLA items. The SLA items exhibited similar function irrespectively of the gender, providing evidence for the validity of the Greek version of the SLA, something that was also found in a similar study with older children (Morgan et al., 2016). This finding is not congruent with a widely accepted perception that boys are likely to perform better in Science Technology Engineering & Mathematics. Yet, there are studies challenging the above notion revealing that there isn't any difference in performance in science knowledge between boys and girls (Mantzicopoulos et al., 2008; Samarapungavan et al., 2009; Guo et al., 2015). Guo et al. (2015) argued that the gender gap in science knowledge maybe a finding for older children and that girls in early childhood do not lag in their competence to learn science. On the other hand, Fleer (1990) argued that the limited exposure to science education in ECE might be the reason for the above result.

We have to be very cautious in reporting firm conclusions, considering the age group under investigation, the relativeness in child evolution at this period of their life, the different instructional guidelines in curricula, the different content and approaches in teaching, and so many other factors that could affect children performance. In her paper, Fleer (1990, p. 366) concluded that in order to have a balance between males and females in their performance in science, "then a concerted effort to conduct research in technology education, commencing at the early childhood level, is urgently needed." This investigation has to be continued further by taking into account additional cultural and societal factors and focused on specific content knowledge each time.

## LIMITATIONS

Although the current study's methodology relied on rigorous statistical methods (IRT, structural equation modeling), it is not without limitations. The current results are encouraging, but other types of validity should also be examined in order for the Greek version of SLA to be established as an instrument of choice to test science knowledge. Criterion and convergent validity together with reliability testing are important aspects of

the SLA's applicability, and further research activities should be adopted toward this direction, before drawing firm conclusions for the instrument's final psychometric properties. Moreover, as the data were cross-sectional it would be interesting to also collect data in different time periods during the academic year, in order to inspect how children's knowledge evolves. Finally, teacher competencies could not be tested in the current study due to limited resources and time. Future studies should take this factor into consideration as it may influence children's knowledge (Gomes and Fleer, 2018; Sundberg et al., 2018).

## CONCLUSIONS

Children in early childhood are qualified learners even for demanding and complex understandings of science learning (Bonawitz et al., 2009; Samarapungavan et al., 2009, 2011; Mantzicopoulos et al., 2013; Zucker et al., 2016). Yet, studies argued that the lack of valid and reliable instruments hinders the efforts of integrating science learning and instruction in ECE (Brenneman, 2011; Kloos et al., 2012; Zucker et al., 2016). To this respect, the assessment of the psychometric properties of an instrument assessing preschoolers' science knowledge seems to contribute toward the expansion of science learning implementation in ECE. The purpose of the current study was to provide to the Greek ECE a valid and reliable instrument that would serve as a tool to assess science learning. The Greek version of the SLA showed adequate psychometric properties regarding the factor structure and item difficulty and discrimination. The shortcomings revealed could be a spark in order to test further the SLA in the field practice. Its application in studies in the USA and Greece revealed that it is a promising evaluation instrument capturing universal concepts of science knowledge children in early childhood, and that it can be used in diverse environments. The provision of a valid instrument for assessing children's science learning, using a strong and flexible

theoretical measurement model (Edwards, 2009) is adding value to the Greek ECE. Having instruments with sound psychometric properties to assess children's learning in various disciplines is a prerequisite for the successful design and implementation of curricula and programs in education field. Otherwise, the utility of assessing children's learning can be considered ambiguous. Thus, in the field of science learning in the Greek ECE, the SLA could help policy makers, researchers, and practitioners to assess effectively children's science learning based on a valid measure.

In order to create strong evidence base for science knowledge in ECE it is essential that researchers have access to valid assessment tools. To this respect, instruments that accurately assess science knowledge in ECE can provide robust evidence about students' learning outcomes. The SLA can serve as an example of such an instrument by providing acceptable psychometric characteristics in the current study. Conclusively, even if the Greek version of the SLA appeared to be a competent instrument for the assessment of children's science knowledge in ECE, further studies with rigorous sampling methods and more complex validity and reliable assessments will be needed before drawing firm conclusions.

## ETHICS STATEMENT

This study was carried out in accordance with the recommendations of the ethical committee of Aristotle University of Thessaloniki and written consent was provided by the parents of the participating children.

## AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, approved it for publication, and had an equal participation in the preparation of the manuscript.

## REFERENCES

Akerson, V. L., Weiland, I., and Fouad, K. E. (2015). "Children's ideas about life science concepts," in Research in Early Childhood Science Education, eds K. C. Thundle and M. Sackes. (New York: NY, Springer), 99–123. doi: 10.1007/978-94-017-9505-0_5

Arnold, D. H., Fisher, P. H., Doctoroff, G. L., and Dobbs, J. (2002). Accelerating math development in Head Start classrooms. J. Educ. Psychol. 94, 762–770. doi: 10.1037/0022-0663.94.4.762

Bagiati, A., Yoon, S. Y., Evangelou, D., and Ngambeki, I. (2010). Engineering curricula in early education: describing the landscape of open resources. Early Childhood Res. Pract. 12 (2). Retrieved from: http://ecrp.uiuc.edu/v12n2/bagiati.html

Bonawitz, E. B., Ferranti, D., Horowitz, A., and Schulz, L. (2009). "The Block Makes It Go": causal language helps toddlers integrate prediction, action, and expectations about contact relations. Proc. Cogn. Sci. Soc. 31, 81–86. Available online at: https://escholarship.org/uc/item/0qm2g1rk

Brenneman, K. (2011). Assessment for preschool science learning and learning environments. Early Childhood Res. Pract. 13. Retrieved from: http://files.eric.ed.gov/fulltext/EJ931225.pdf

Brenneman, K., and Louro, I. F. (2008). Science journals in the preschool classroom. Early Childhood Educ. J. 36, 113–119. doi: 10.1007/s10643-008-0258-z

Cai, L., Thissen, D., and du Toit, S. H. C. (2011). IRTPRO for Windows [Computer software and manual]. Lincolnwood, IL: Scientific Software International.

Clements, D. H. (2001). Mathematics in the preschool. Teach. Child. Mathemat. 7, 270–275. Available online at: https://search.proquest.com/docview/214137279/155E375C07104B0EPQ/3?accountid=8359

Clements, D. H., and Sarama, J. (2016). Math, science, and technology in the early grades. Fut. Children 26, 75–94. doi: 10.1353/foc.2016.0013

Connor, C. M., Morrison, F. J., and Petrella, J. (2004). Effective reading comprehension instruction: examining child instruction interactions. J. Educ. Psychol. 96, 682–669. doi: 10.1037/0022-0663.96.4.682

DeMars, C. (2010). Item Response Theory. Oxford, UK: Oxford University Press.

Duncan, G. J., Claessens, A., Huston, A. C., Pagani, L. S., Engel, M., Sexton, H., et al. (2007). School readiness and later achievement. Dev. Psychol. 43, 1428–1446. doi: 10.1037/0012-1649.43.6.1428

Edwards, M. C. (2009). An introduction to item response theory using the need for cognition scale. Social Personal. Compass 3, 507–529. doi: 10.1111/j.1751-9004.2009.00194.x

Embretson, S. E., and Reise, S. P. (2000). Item Response Theory for Psychologists. Mahwah, NJ: Lawrence Erlabaum Associates.

Ergazaki, M., Valanidou, E., Kasimati, M.-C., and Kalantzi, M. (2015). Introducing a precursor model of inheritance to young children. Int. J. Sci. Educ. 37, 3118–3142. doi: 10.1080/09500693.2015.1121551

Eshach, H., and Fried, M. (2005). Should science be taught in early childhood? *J. Sci. Educ. Technol.* 14, 315–336. doi: 10.1007/s10956-005-7198-9

Finch, W. H., and French, B. F. (2007). Detection of crossing differential item functioning: a comparison of four methods. *Educ. Psychol. Meas.* 67, 565–582. doi: 10.1177/0013164406296975

Fleer, M. (1990). Gender issues in early childhood science and technology education in Australia. *Int. J. Sci. Educ.* 12, 355–367. doi: 10.1080/0950069900120403

Fragkiadaki, G., Fleer, M., and Ravanis, K. (2017). A cultural-historical study of the development of Children's scientific thinking about clouds in everyday life. *Res. Sci. Educ.* 1, 1–23. doi: 10.1007/s11165-017-9665-8

French, L. A. (2004). Science as the center of a coherent, integrated early childhood curriculum. *Early Child. Res. Q.* 19, 138–149. doi: 10.1016/j.ecresq.2004.01.004

French, L. A., and Woodring, S. (2012). "Science education in the early years," in *Handbook of Research on the Education of Young Children,* 3rd ed, eds B. Spodek and O. Saracho (New York, NY: Taylor and Francis), 179–185.

Gelman, R., and Brenneman, K. (2004). Science learning pathways for young children. *Early Child. Res. Q.* 19, 150–158. doi: 10.1016/j.ecresq.2004.01.009

Giallousi, M., Tselfes, V., and Gialamas, V. (2014). Using student-teachers' reports of self-efficacy to evaluate an early childhood science course. *J. Early Childhood Teac. Educ.* 35, 337–356. doi: 10.1080/10901027.2014.968298

Ginsburg, H. P., and Golbeck, S. L. (2004). Thoughts on the future of research on mathematics and science learning and education. *Early Child. Res. Q.* 19, 190–200. doi: 10.1016/j.ecresq.2004.01.013

Gomes, J., and Fleer, M. (2018). Is science really everywhere? Teachers' perspectives on science learning possibilities in the preschool environment. *Res. Sci. Educ.* doi: 10.1007/s11165-018-9760-5. [Epub ahead of print].

Grammatikopoulos, V. (2012). Integrating program theory and systems based procedures in program evaluation: a dynamic approach to evaluate educational programs. *Educ. Res. Eval.* 18, 53–64. doi: 10.1080/13803611.2011.640874

Grammatikopoulos, V., Linardakis, M., Gregoriadis, A., and Oikonomidis, V. (2015). Assessing the students' evaluations of educational quality (SEEQ) questionnaire in Greek higher education. *Higher Educ.* 70, 395–408. doi: 10.1007/s10734-014-9837-7

Grammatikopoulos, V., Zachopoulou, E., Tsangaridou, N., Liukkonen, J., and Pickup, I. (2008). Applying a mixed method design to evaluate training seminars within an early childhood education project. *Eval. Res. Educ.* 21, 4–17. doi: 10.2167/eri417.0

Greenfield, D. B. (2015). "Assessment in early childhood science education," in *Research in Early Childhood Science Education*, eds K. C. Thundle and M. Sackes (New York, NY: Springer), 353–380. doi: 10.1007/978-94-017-9505-0_16

Guo, Y., Piasta, S. B., and Bowles, R. P. (2015). Exploring preschool children's science content knowledge. *Early Educ. Dev.* 26, 125–146 doi: 10.1080/10409289.2015.968240

Hammer, A. S. E., and He, M. (2014). Preschool teachers' approaches to science: a comparison of a Chinese and a Norwegian kindergarten. *Eur. Early Childhood Educ. Res. J.* 24, 450–464. doi: 10.1080/1350293X.2014.970850

Hamre, B. K., and Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk for school failure? *Child Dev.* 76, 949–967. doi: 10.1111/j.1467-8624.2005.00889.x

Hartig, J., and Höhler, J. (2008). Representation of competencies in multidimensional IRT models with within-item and between-item dimensionality. *J. Psychol.* 216, 89–101. doi: 10.1027/0044-3409.216.2.89

Hu, L. T., and Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Struct. Equat. Model.* 6, 1–55. doi: 10.1080/10705519909540118

Kallery, M. (2004). Early years teachers' late concerns and perceived needs in science: an exploratory study. *Eur. J. Teac. Educ.* 27, 147–165. doi: 10.1080/026197604200023024

Kallery, M., Psillos, D., and Tselfes, V. (2009). Typical didactical activities in the greek early years science classroom: do they promote science learning? *Int. J. Sci. Educ.* 31, 1187–1204. doi: 10.1080/09500690701824850

Kalogiannakis, M., Nirgianaki, G., and Papadakis, S. (2018). Teaching magnetism to preschool children: The effectiveness of picture story reading. *Early Childhood Educat. J.* 46, 535–546. doi: 10.1007/s10643-017-0884-4

Kalogiannakis, M., and Papadakis, S. (2019). Evaluating pre-service kindergarten teachers' intention to adopt and use tablets into teaching practice for natural sciences. *Int. J. Mobile Learn. Organ.* 13, 113–127. doi: 10.1504/IJMLO.2019.096479

Kanaki, K., and Kalogiannakis, M. (2018). Introducing fundamental object-oriented programming concepts in preschool education within the context of physical science courses. *Educ. Inform. Tech.* 23, 2673–2698. doi: 10.1007/s10639-018-9736-0

Kelley, T. L. (1939). "e selection of upper and lower groups for the Validation of test items. *J. Educ. Psychol.* 30, 17–24. doi: 10.1037/h0057123

Kloos, H., Baker, H., Luken, E., Brown, R., Pfeiffer, D., and Carr, V. (2012). *Preschoolers Learning Science: Myth or Reality*? INTECH, Open Science, Open Minds. Retrieved from: http://cdn.intechopen.com/pdfs/40973/InTech-Preschoolers_learning_science_myth_or_reality_.pdf

Mantzicopoulos, P., Patrick, H., and Samarapungavan, A. (2008). Young children's motivational beliefs about learning science. *Early Child. Res. Q.* 23, 378–394. doi: 10.1016/j.ecresq.2008.04.001

Mantzicopoulos, P., Patrick, H., and Samarapungavan, A. (2013). Science literacy in school and home contexts: kindergarteners' science achievement and motivation. *Cogn. Instr.* 31, 62–119. doi: 10.1080/07370008.2012.742087

Mantzicopoulos, P., Samarapungavan, A., and Patrick, H. (2009). We learn how to predict and be a scientist": early science experiences and kindergarten Children's social meanings about science. *Cogn. Instr.* 27, 312–369. doi: 10.1080/07370000903221726

Metz, K. (2004). Children's understanding of scientific inquiry: their conceptualization of uncertainty in investigations of their own design. *Cogn. Instr.* 22, 219–290. doi: 10.1207/s1532690xci2202_3

MoE/PI, I. (2002). *Cross-Thematic Curriculum Framework for Compulsory Education (DEPPS)*. Athens: Ministry of National Education and Religion.

Morgan, P. L., Farkas, G., Hillemeier, M. M., and Maczuga, S. (2016). Science achievement gaps begin very early, persist, and are largely explained by modifiable factors. *Educ. Res.* 45, 18–35. doi: 10.3102/0013189X16633182

Muthen, L. K., and Muthen, B. (2012). 1998–2012. *Mplus User's Guide*, 7th Edn. Los Angeles, CA: Muthen & Muthen.

Orlando, M., and Thissen, D. (2003). Further investigation of the performance of *S-X2*: An item fit index for use with dichotomous item response theory models. *Appl. Psychol. Meas.* 27, 289–298. doi: 10.1177/0146621603027004004

Penfield, R. D., and Camilli, G. (2007). "Differential item functioning and item bias," in *Handbook of Statistics* Vol. 26, eds S. Sinharay and C. R. Rao (New York, NY: Elsevier), 125–167. doi: 10.1016/S0169-7161(06)26005-X

Samarapungavan, A., Mantzicopoulos, P., and Patrick, H. (2008). Learning science through inquiry in kindergarten. *Sci. Educ.* 92, 868–908. doi: 10.1002/sce.20275

Samarapungavan, A., Mantzicopoulos, P., Patrick, H., and French, B. F. (2009). The development and validation of the Science Learning Assessment (SLA): a measure of kindergarten science learning. *J. Adv. Acad.* 20, 502–535. doi: 10.1177/1932202X0902000306

Samarapungavan, A., Patrick, H., and Mantzicopoulos, P. (2011).What kindergarten students learn in inquiry-based science classrooms. *Cogn. Instr.* 29, 416–470. doi: 10.1080/07370008.2011.608027

Singh, J. (2004). Tackling measurement problems with Item Response Theory. Principle, characteristics, and assessment with an illustrative example. *J. Business Res.* 57, 184–208. doi: 10.1016/S0148-2963(01)00302-2

Sundberg, B., Areljung, S., Due, K., Ekström, K., Ottander, C., and Tellgren, B. (2018). Opportunities for and obstacles to science in preschools: views from a community perspective. *Int. J. Sci. Educ.* 40, 2061–2077. doi: 10.1080/09500693.2018.1518615

Tsigilis, N., Grammatikopoulos, V., and Koustelios, A. (2007). Applicability of the Teachers' Sense of Efficacy Scale to educators teaching innovative programs. *Int. J. Educ. Manag.* 21, 634–642. doi: 10.1108/09513540710822229

Vitiello, V. E., Whittaker, J. V., Mulcahy, C., Kinzie, M. B., and Helferstay, L. (2019). Reliability and validity of the preschool science observation measure. *Early Educ. Dev.* 30, 196–215. doi: 10.1080/10409289.2018.1544814

Wang, J., and Wang, X. (2012). *Structural Equation Modeling. Applications Using Mplus*. Beijing: Higher Education Press.

Watters, J. J., Diezmann, C. M., Grieshaber, S. J., and Davis, J. M. (2001). Enhancing science education for young children: a contemporary initiative. *Aust. J. Early Child.* 26, 1–7. doi: 10.1177/183693910102600202

Yu, C.-Y. (2002). *Evaluating Cutoff Criteria of Model Fit Indices for Latent Variable Models with Binary and Continuous Outcomes.* Los Angeles, CA: Unpublished doctoral dissertation University of California.

Zhao, J., and Gallant, D. J. (2012). Student evaluation of instruction in higher education: exploring issues of validity and reliability. *Assess. Eval. Higher Educ.* 37, 227–235. doi: 10.1080/02602938.2010.523819

Zucker, T. A., Williams, J. M., Bell, E. R., Assel, M. A., Landry, S. H., Monsegue-Bailey, P., et al. (2016). Validation of a brief, screening measure of low-income pre-kindergarteners' science and engineering knowledge. *Early Child. Res. Q.* 36, 345–357. doi: 10.1016/j.ecresq.2015.12.018