Journal of Mathematics and Science Teacher
2023, 3(1), em031
e-ISSN: 2752-6054
https://www.mathsciteacher.com/

MODESTUM

Research Article **OPEN ACCESS**

The evaluation of GAENE 2.1 in Greek students

Miltiadis Manikas [1*], Ioannis Leonardos [1], Katerina Plakitsi [2], Konstantinos Sotiropoulos [1]

[1] Department of Biological Applications & Technology, University of Ioannina, GREECE
[2] Department of Early Childhood Education, University of Ioannina, GREECE
***Corresponding Author:** mmanik71@gmail.com

Citation: Manikas, M., Leonardos, I., Plakitsi, K., & Sotiropoulos, K. (2023). The evaluation of GAENE 2.1 in Greek students. *Journal of Mathematics and Science Teacher, 3*(1), em031. https://doi.org/10.29333/mathsciteacher/13069

ARTICLE INFO

Received: 26 Feb. 2023
Accepted: 08 Mar. 2023

ABSTRACT

The acceptance of evolution by Greek students aged 17 has never been actually measured neither via a poll survey nor with the application of any of the existing psychometric tools that have been developed for this purpose. In this research, GAENE 2.1 was chosen because it is a one-dimensional tool that can only measure the acceptance of evolution and does not contain items related to the understanding of evolution and the respondent's religious beliefs. The assessment of the appropriateness of the tool for our sample was tested with the criteria of Rasch model. Despite the fact that some of GAENE 2.1 items may not show the expected statistical fit and probably overestimates the levels of acceptance of evolution, the tool is appropriated to measure with efficiently and effectively way the acceptance of evolution in the specific sample.

Keywords: acceptance of evolution, GAENE 2.1, Rasch analysis

INTRODUCTION

Evolution is the framework in which many biological phenomena are interpreted. These phenomena have not only received a lot of research attention by scientists, but also affect the daily life of people as they are related to the emergence of new infectious diseases, bacterial resistance to antibiotics, insect resistance to insecticides etc. (Dunk et al., 2019). For this purpose, in many curricula across the world the role have the evolution has been viewed as a core theme in science education (AAAS, 2011; Eder et al., 2018; Harms & Reiss, 2019; Kazempour & Amirshokoohi, 2018; NGSS Lead States, 2013; Quessada & Clement, 2018).

The theory of evolution is one of the most important and robust scientific theories nowadays (Sfendourakis, 2021). Evolutionary theory is the unifying theory of biology (Dobzhansky, 1973) that makes biology independent discipline of science (Lennox, 1992). At the same time it is applied in non-biological fields too (Harms & Reiss, 2019). However, despite its universality and its significance for the daily life of people, is generally overlooked thus a number of misunderstandings and misinterpretations have occurred (Dawkins, 2009; Gregory, 2009; Kampourakis & Zogza, 2008; Prinou et al., 2008). In addition, to date it is more common for the general public to frame evolution within supernatural and religious beliefs (Gallup, 2019; Miller et al., 2006).

General public acceptance of evolution has received a lot of research attention. Surveys that have been conducted for this purpose so far usually include a few items. As a result, these measures have limited credibility (Smith et al., 2016) and their conclusions can be misleading (Kuschmierz et al., 2021). In contrast, psychometric instruments have been developed, and they have been applied in various studies and samples, mainly on students and teachers. These instruments usually consist of several items, which should measure only one theoretical construction–the acceptance of evolution (AERA, 2014; Boone et al., 2014) and according to the researchers have a high degree of reliability and validity (Smith et al., 2016).

The most widely used instrument for measuring the acceptance of evolution is the measure of acceptance of the theory of evolution (MATE) (Mead et al., 2019). MATE was originally developed to investigate the levels of acceptance of evolutionary theory by high school biologists (Rutledge & Warder, 1999). Later, its application was extended to other population groups, mainly to students at various university departments (Rutledge & Salder, 2007). However, research data show that its validity is not clear (Smith, 2010; Smith et al., 2016) and at the same time there are vague indications of its uni-dimensionality (Metzger et al., 2018; Romine et al., 2017; Sbeglia & Nehm, 2018). It also discriminates between five subgroups of acceptance of evolution without presenting data on how the boundaries of each category are set (Smith et al., 2016).

To the best of our knowledge, MATE is the only instrument that has been used to measure the acceptance of evolutionary theory in Greek students and teachers (Athanasiou et al., 2012, 2016; Athanasiou & Papadopoulou, 2012, 2015; Katakos & Athanasiou, 2020; Mantelas & Mavrikaki, 2020; Stasinakis & Athanasiou, 2016), without any evaluation of its application in that specific samples. At the same time, neither this nor any other instrument for the acceptance of the evolution has ever been applied to secondary school students.

To date, various psychometric tools (EALS, I-SEA, and GAENE) have been developed that improve the ability to measure the acceptance of evolution. They are based on

(a) the evaluation of the items by experts and not just in one round,

(b) the implementation of a large number of participants in the initial evaluation,

(c) the clear definition of the concept of acceptance, and

(d) the application of various statistical models for the evaluation of the instrument both as a whole and each item separately (Smith et al., 2016).

In many cases, the role of Rasch measurement model being a uni-dimensional one, has been viewed as an instrument measuring only one variable or trait (Boone & Noltemeyer, 2017; Tennant & Conaghan, 2007). Rasch model can provide results for the reliability and validity of an instrument by utilizing several indicators and produce objective measurements for parametric statistical analyses by converting the Likert scale into an equal interval scale (Boone et al., 2014). Simultaneously, it can evaluate the quality of the rating scale of an instrument with many criteria (Fisher, 2007).

GAENE 2.1 (Smith et al., 2016), like other psychometric tests, uses the Likert scale to express the degree of agreement or disagreement in each of the measurement tool items. The Likert scale includes categories such as “strongly disagree” (SD), “disagree” (D), “I’m uncertain /cannot decide” (U), “agree” (A) and “strongly agree” (SA), which correspond to the numbers 1-5 or vice versa and then treated as real numbers rather than as a form of convention that allows data to be coded (Wu & Leung, 2017). These labels are used incorrectly in descriptive and parametric statistical analyses because the consecutive intervals between them are unequal (Stevens, 1946). Thus, they can produce distorted results (Planinic et al., 2019), which imply further speculations (Törmäkangas, 2011). Matching Likert scale categories with numbers does not automatically convert them into equal intervals (Knapp, 1990). Also, in many surveys, reference is made to the average of the respondents’ responses to an item of a tool in the form of, for example, 4.23. Recoding this average reveals that 4.23=I agree.23 (Kuzon et al., 1996). Thus, Rasch model was chosen for the analysis of the data.

According to the former model, when an attempt is made to measure a variable from a psychometric tool that includes several items, then the people who complete the questionnaire will be much more likely to answer the easy ones than the most difficult. At the same time, people with a large ability with respect to the variable measured by the tool will be more likely to respond correctly to all the items than people with low ability (Boone & Noltemeyer, 2017). This technique allows Rasch model to calculate linear personal measurements from raw data (Boone, 2016; Metzger et al., 2018) and to use these measurements to perform reliable statistical analyses (Schumacker, 2016).

Also, if the data were fitted to Rasch model then the estimation for the position of the items is independent of the distribution of the individual measurements and vice versa (Metzger et al., 2018; Smith et al., 2002). This estimation between the person and items of the tool is not possible when the raw data is used, because the ability of the persons to respond varies according to the items of the tool and the items depend on the person who has completed the questionnaire (Smith et al., 2002). In previous studies with GAENE 2.1 in the US has shown that the items of the instrument fit well Rasch model (Metzer et al., 2018; Sbeglia & Nehm, 2018).

MATERIALS AND METHODS

This study sought to evaluate GAENE 2.1 as a tool to measure the acceptance of evolution in 17 year old Greek students attending state senior high school. More on this we are also examined the levels of the acceptance of evolution in that age group.

Although relatively past research conducted on a similar sample of students had tended to show a positive view of the evolution, it did not apply a published instrument to measure the levels of acceptance of evolution. Furthermore, no data established for the validity and reliability of the items that are used for having drawn such a conclusion (Prinou et al., 2008).

The present research will explore the following questions:

1. **RQ1.** Is GAENE 2.1 an appropriate instrument for measuring the acceptance of the evolution among Greek students aged 17 attended the 2nd grade of state senior high school and have not yet been taught evolutionary theory?
2. **RQ2.** What are the levels of acceptance of evolution among students in this age group and how are they influenced by the gender and the field of studies chosen by the student?

Materials

Participants completed a questionnaire of two parts in order to gain insight into the former questions. The first part includes general information, such as gender, the direction they are attending, the field of studies they are thinking of following in the next class, etc. The second part consists of the 13 items (1-13) of GAENE 2.1 (Smith et al., 2016). The items were translated into Greek and then re-translated into English, to ensure the validity of the translation (**Table 1**).

The statistical analysis of the questionnaires was made with the help of the statistical software SPSS 25 and the WINSTEPS (version 5.2.2.0), (Linacre, 2022). The data were entered for descriptive and parametric statistical analyses in SPSS 25. The initial analysis for the evaluation of GAENE 2.1 was done with the help of WINSTEPS, which it supports and performs statistical analyses of Rasch model.

Table 1. The GAENE 2.1. items (Smith et al., 2016)

a/a	Item
1	Everyone should understand evolution.
2	It is important to let people know about how strong the evidence that supports evolution is.
3	Some parts of evolution theory could be true.
4	Evolutionary theory applies to all plants and animals, including humans.
5	People who plan to become biologists need to understand evolution.
6	I would be willing to argue in favor of evolution[a] in a public forum such as a school club, church group, or meeting of public school parents.
7	Simple organisms such as bacteria change over time.
8	Nothing in biology makes sense without evolution.
9	Understanding evolution helps me understand the other parts of biology.
10	I would be willing to argue in favor of evolution in a small group of friends.
11	Evolution is a good explanation of how humans first emerged on the earth.
12	Evolution is a scientific fact.
13	Evolution is a good explanation of how new species arise.

Note. [a]In this item, Smith et al. (2016) used the word "evolutionary", which according to Sbeglia and Nehm (2018) it is a typo. Thus, I follow the proposition of Sbeglia and Nehm (2018) and I modified it to "evolution" in here and in the questionnaire administered to students.

Participants

The study was conducted in March 2021 and its sample consisted of 117 students (61 girls and 56 boys) of the B class of two state senior high school, urban area of the geographical region of Epirus in Greece. The total number of students who should have participated in the survey was 121, but for various reasons when the questionnaire was administrated, four students were absent from school. The participants in this study were categorized based on the field of studies that they would follow in the next class. So, of the 117 students, 23 will choose the humanities studies, 34 will choose the field of health studies, 34 the field of science studies and the rest 26 the field of economics and computer studies.

Also, the students who participated in the present study have not yet been taught evolutionary theory in senior high school. However, 51 of the 117 have taken evolutionary courses during their C grade of junior high school while 66 have not been taught anything about the evolution of the living world.

Methods

The raw data were entered into the statistical software SPSS 25 and both nominal and ordinal variables were numerically coded. Each item of the GAENE 2.1 was coded with the numbers 1 (strongly disagree) to 5 (strongly agree). Thus, people with low acceptance in evolution will also have low performance in raw data and vice versa. Also, because the 5-point scale was chosen for the research–as suggested by the researchers who developed the tool (Smith et al., 2016)–the minimum performance that one can achieve is 13 and the maximum is 65.

Then the data were entered into the WINSTEPS and several properties of the instrument for the specific sample were examined, in order to answer the research questions, such as:

(a) its dimensionality,

(b) the fit of the items to Rasch model,

(c) its operation of 5-point scale,

(d) its reliability,

(e) its construct validity, and

(f) the distribution of the difficulty of the items and the responses of the persons (Wright map).

These properties will be used as criteria for evaluating the quality and capacity of the tool to provide reliable results on the levels of acceptance of evolution in this sample.

Finally, with the help of the SPSS 25 and the use of parametric statistical analyses, the gender, the field of study that they will follow in the final year of the senior high school and the previous courses in the evolutionary theory will be assessed as factors influencing the acceptance of evolution. Simultaneously, the average performance of the students per item of the tool and the overall result during the application of GAENE in the specific sample will be evaluated, so that it is possible to compare the results with other surveys that have been conducted with the specific tool (Sbeglia & Nehm, 2018).

The instruments that used to estimate a trait or a theoretical construct, such as acceptance of evolution, should be unidimensional (Hagell, 2014). This means that all items of the instrument should measure a trait and each of them should measure a slightly different aspect of that trait (Andrich & Marais, 2019). The uni-dimensionality of the instrument is a prerequisite for the application of Rasch model to the data collected and then used to calculate measurements for persons or for items (Boone et al., 2014).

One important element for the uni-dimensionality of an instrument is the principal components analysis of the residuals (Tennant & Conagham, 2007; Sbeglia & Nehm, 2018). Rasch model predicts that if there are components of unexplained variance with eigenvalues greater than 2.00, may be evidence of a second major dimension in the tool (Linacre, 2021; Sbeglia & Nehm, 2018). Also, if the first component of unexplained variance has a higher value than the components of the variance that explained of the data, may be an indicator of a second major dimension in the tool, too (Linacre, 2021).

Another element for the uni-dimensionality of an instrument is the average MNSQ values of the ITEM: INFIT & OUTFIT indices and the MNSQ values for each item separately. If the average MNSQ for the items is close to 1.00 then it is an element in favor of the uni-dimensionality (Arnold et al., 2018; Wright & Linacre, 1994). In order to evaluate the way, which the items fit Rasch model for a particular sample, the indices infit and outfit (MNSQ and ZSTD) are calculated. The indicators infit (MNSQ) and outfit (MNSQ) are chi square statistics and the value expected for them is 1.0. Values in these indices higher than 1.00 indicate that there is too much unexplained variance in the data and values lower than 1.00 indicate that the data are too predictable for model (Boone et al., 2014). Infit (ZSTD) and outfit (ZSTD) are actually t-tests that indicate whether the values of the MNSQ indices are due to the random factors (Boone et al., 2014). Items using the Likert scale have a range of good fit of the MNSQ indices between the values 0.6-1.4 and the ZSTD indices from -1.9 to +1.9 (Wright & Linacre, 1994). Although some researchers suggest the range between 0.5-1.5 as acceptable limits for MNSQ indices (Boone et al., 2014; Metzger et al., 2018). An item does not fit well into the model if the participants who seem to bear low ability in responding accurately to the variable, agree with a difficult item and vice versa (Boone et al., 2014).

Rasch model provides the ability to evaluate the operation of the categories of a psychometric tool. The categories of the scale work satisfactory when

(a) each category seems to have the highest probability of appearing in a specific combination of difference between individual measurement and the difficulty of items (Boone et al., 2014) and

(b) the Andrich thresholds are not too close one another and for each item of the tool a normal distribution is displayed (Sbeglia & Nehm, 2018).

Thus, the separation between the hills at each of the categories of the scale and the normal distribution of each item indicates the ability of the individuals can be distinguished from the instrument items in the way it is predicted by Rasch model (Smith et al., 2016; Tennant & Conaghan, 2007).

The reliability of an instrument is related to the stability of the measurement (Heale & Twycross, 2015) and although, it is difficult to test it in educational research (Taber, 2018), it is estimated by using Cronbach's alpha index. Cronbach's alpha index could be calculated when all items of the instrument measure a single construction (Gardner, 1995; Taber, 2018) and provides indications about the operation of the items of the instrument (Taber, 2018). However, the use of raw data for the calculation of this index can lead to an overestimation of the reliability of the tool since the categories of the scale are not equal intervals (Linacre, 1997, 2002). In Rasch model the value corresponding to the Cronbach's alpha index is the PERSON RELIABILITY. The value of this index depends on various factors, which do not include the specific items of the instrument (Cronbach, 1951), but others, such as

(a) the range of variation in the sample,

(b) the size of the instrument,

(c) the number of categories on the Likert scale, etc. (Boone et al., 2014; Linacre, 2021).

The PERSON RELIABILITY indicates the ability of an instrument, which measures a theoretical construct, to separate the person in relation to the variable (Linacre, 2021) and evaluates the probability of replicability of the results for the same sample (Arnold et al., 2018).

The reliability of the items is measured by Rasch model with a different index, the ITEM RELIABILITY. The value of this index depends on the size of the sample and the variability of the difficulty of the items (Boone et al., 2014; Linacre, 2021) and provides indications of the replicability of the item hierarchy when the instrument is applied to a different sample (Arnold et al., 2018; Linacre, 2021).

The validity is defined as the degree to which a quantitative survey can measure a variable (Heale & Twycross, 2015) and allows conclusions about the variable to be drawn from the sample (Creswell, 2012). Usually content validity is checked by evaluating the items that make up the tool by a team of experts at one or more rounds (Smith et al., 2016).

The construct validity can be assessed by the behavior of the items in terms of difficulty and how they are distributed on the Wright map. In Rasch model the average difficulty of the items of an instrument is set to zero (Boone et al., 2014; Tennant & Conagham, 2007) and they should be evenly distributed from the lower to the upper end of the measurements, while at the same time there should be no overlap, i.e., there should be no items in the same level of difficulty (Boone et al., 2014). Another element for the validity of an instrument it could be provided by the comparison of the hierarchy of the difficulty of the items with data available in the literature (Arnold et al., 2018; Boone et al., 2014).

The difficulty of the items and the ability of the students to answer each item are presented with the Wright map. This representation shows the relationship between the individual measurements and the measurements of the items in the same unit (the logit). Thus, we ascertain that people with low acceptance of evolution and items with which one can easily agree are at the bottom of the map. On the contrary, people with high acceptance of evolution and items, which one can hardly agree with are at the top of the map (Boone et al., 2014).

Wright map provides indications about the quality of targeting of persons and items because it is easy to compare the average of the person measures with the average of the difficulty of the tool items (Tennant & Conagham, 2007). When the average of the person measures is less than zero, it means that the items of the tool are difficult to agree with them and vice versa (Boone et al., 2014).

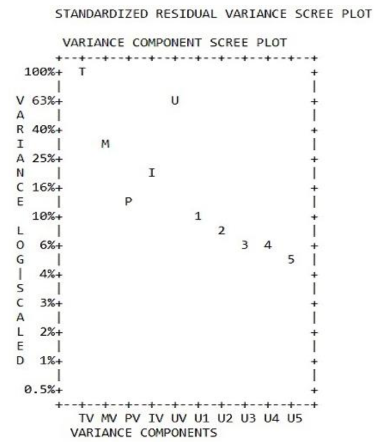

Figure 1. Principal component analysis of standardized residuals (Winsteps: Rasch Measurement Computer Program)

Table 2. Variance of standardized residual

a/a	Sort of variance	Symbol on plot in Figure 4 (& in x axis)	Eigenvalue	% variance from data	% variance expected from Rasch model
1	Total raw variance (T=M+U)	T (TV)	20.26	100.0%	100.0%
2	Raw variance explained by measures (M=P+I)	M (MV)	7.26	35.8%	36.0%
3	Raw variance explained by persons (P)	P (PV)	3.18	15.7%	15.7%
4	Raw Variance explained by items (I)	I (IV)	4.08	20.2%	20.2%
5	Raw unexplained variance (U=U1+U2+ …)	U (UV)	13.00	64.2%	64.0%
6	Unexplained variancein first contrast (U1)	1 (U1)	2.08	10.3%	16.0%
7	Unexplained variancein second contrast (U2)	2 (U2)	1.64	8.1%	12.6%

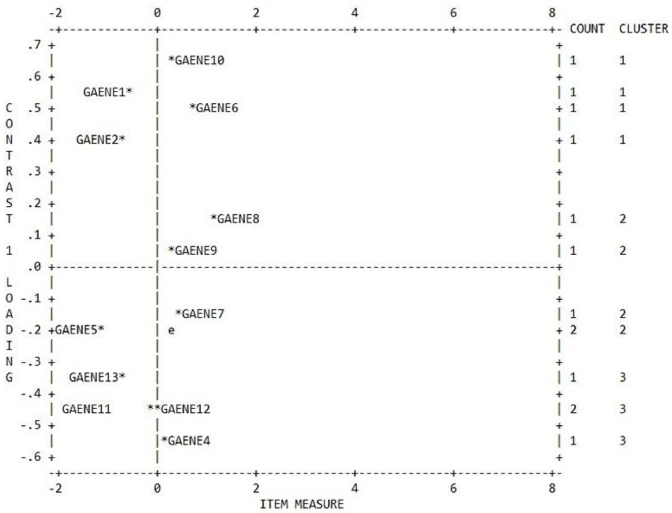

Figure 2. Clusters of items according to the first contrast of residuals (Winsteps: Rasch Measurement Computer Program)

RESULTS

RQ1

The dimensionality analysis (**Figure 1**) performed on the sample data showed that the percentage of variance explained by the measurements was M=35.8%. This percentage is explained by the variance that comes from the person measures P=15.7% and by the variance that comes from the items of the instrument I=20.2%. Although the percentage of unexplained variance is high UV=64.2%, it is not a problem for the one-dimensionality of the instrument in the specific sample, because the percentage of variance from the first contrast of the residuals is less than the percentage of variance from the person measures as well as the percentage of variation that comes from the items of the instrument (U1=10.3%) (Linacre, 2021).

However, the eigenvalue of the first contrast (**Table 2**) is marginally greater than two (2.08), which may be an indication of the existence of a second dimension in the data (Linacre, 2021; Sbeglia & Nehm, 2018). Still, in some cases such values may be firstly due to random factors (Raiche, 2005), or to the homogeneity of the sample as students came from only two schools, or to the distribution of items of the instrument in relation to their difficulty (Linacre, 2021).

Figure 2 shows the clusters of items as they are classified by the loadings in the first contrast of the residuals.

Table 3. Disattenuated corelletion between item clusters

PCA contrast	Item clusters	Disattenuated correlation
1	1-3	0.65
1	1-2	0.68
1	2-3	0.83
2	1-3	0.51
2	1-2	0.71
2	2-3	0.93
3	1-3	0.72
3	1-2	0.75
3	2-3	0.80

Table 4. Measures of items GAENE 2.1.

Entry No	Total score	Mean by item	Total count	Measure	Model SE	Infit		Outfit		Item
						MNSQ	ZSTD	MNSQ	ZSTD	
1	474	4.05	117	-0.56	0.15	0.77	-1.75	0.76	-1.91	GAENE 1
2	478	4.09	117	-0.66	0.16	0.83	-1.25	0.84	-1.20	GAENE 2
3	430	3.74	115	+0.23	0.14	0.82	-1.37	0.85	-1.18	GAENE 3
4	434	3.77	115	+0.15	0.15	0.97	-0.20	0.99	-0.04	GAENE 4
5	484	4.24	114	-1.16	0.16	1.19	1.35	1.10	0.74	GAENE 5
6	410	3.50	117	+0.77	0.14	1.25	1.84	1.22	1.65	GAENE 6
7	428	3.66	117	+0.43	0.14	1.05	0.45	1.20	1.48	GAENE 7
8	385	3.32	116	+1.17	0.13	1.25	1.85	1.27	1.97	GAENE 8
9	417	3.69	113	+0.33	0.14	1.09	0.67	1.11	0.82	GAENE 9
10	429	3.70	116	+0.34	0.14	0.91	-0.67	0.88	-0.91	GAENE 10
11	445	3.90	114	-0.17	0.15	0.97	-0.17	0.95	-0.30	GAENE 11
12	446	3.88	115	-0.10	0.15	1.06	0.45	1.07	0.57	GAENE 12
13	479	4.13	116	-0.78	0.16	0.72	-2.17	0.74	-2.05	GAENE 13
Mean	441.5	3.82	115.5	0.00	0.15	0.99	-0.07	1.00	-0.03	
PSD	29.0	0.61	1.3	0.63	0.01	0.17	1.28	0.17	1.30	

Note. The column Entry No corresponds to the item of the instrument; the column Total score corresponds to the sum of the raw data for the specific item; the column Measure is the value that calculated from the application of Rasch model on the raw data; & the columns Infit & outfit are indices for the fit of items on Rasch model

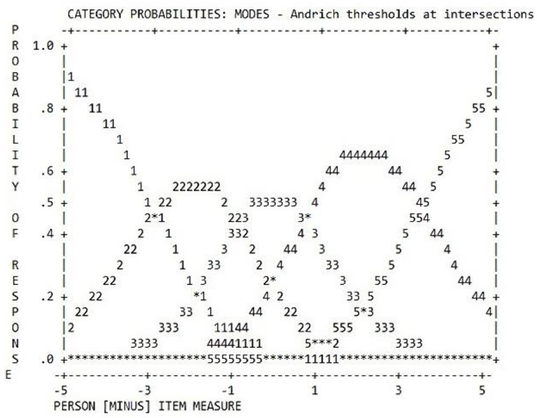

Figure 3. Probability curves for each rating scale category (dotted lines represent the probability of appearance of a particular combination between person measures and item difficulty; dashed lines show Andrich thresholds for each pair of categories [1=SD, 2=D, 3=U, 4=A, 5=SA]). (Winsteps: Rasch Measurement Computer Program)

The disattenuated correlations between the three clusters range from 0.65-0.83 (**Table 3**), i.e., they are above the threshold (<0.57), which indicates the clear existence of a second dimension in the data (Linacre, 2021). Thus, GAENE is a one-dimensional tool for this sample, as the research that have been conducted with it in the USA so far have shown (Metzger et al., 2018; Sbeglia & Nehm, 2018; Smith et al., 2016).

The fit indices of the thirteen elements of the instrument have an average close to 1.00 (INFIT: MNSQ=0.99 and OUTFIT: MNSQ=1.00) constituting an additional element for the one-dimensionality of the tool (Arnold et al., 2018). Each element of the tool has values of the above indicators ranging from 0.72-1.27 (**Table 4**). The fit of the items of the instrument to Rasch model, using the 5-point scale, is very good for all of them (**Table 4**). The OUTFIT MNSQ index, which is the most sensitive (Boone et al., 2014), ranges from 0.74 to 1.27 (Fisher, 2007) and only two items (GAENE 1 and GAENE 13) are out of strict limits.

The effectiveness of the 5-point scale to separate in the way predicted by the model, the responses of students with different levels of acceptance of evolution, is shown in **Figure 3** (Smith et al., 2016).

Table 5. Person & item reliability

A: Person reliability				
	Total score	**Count**	**Measure**	**Model SE**
Mean	49.1	12.8	1.68	0.45
Max	62.0	13	4.73	0.66
Min	35.0	10	-0.44	0.38
Real	RMSE=0.49	True SD=0.91	Separation=1.85	Person reliability=0.77
Model	RMSE=0.45	True SD=0.94	Separation=2.08	Person reliability=0.81
SE of person mean=0.10 & Cronbach's alpha=0.81			SEM=2.39	
B: Item reliability				
	Total score	**Count**	**Measure**	**Model SE**
Mean	441.5	115.5	0.00	0.15
Max	484	117	1.17	0.16
Min	385	113	-1.16	0.13
Real	RMSE=0.15	True SD=0.61	Separation=4.02	Item reliability=0.94
Model	RMSE=0.15	True SD=0.61	Separation=4.16	Item reliability=0.95
SE of item mean=0.18				

Table 6. Ranking of GAENE 2.1 items in various studies

Measure order	**Smith et al. (2016)**		**Romine et al. (2018)**		**Sbeglia and Nehm (2018)**		**This study**	
	Entry number of items	**Difference according to the rank**	**Entry number of items**	**Difference according to the rank**	**Entry number of items**	**Difference according to the rank**	**Entry number of items**	**Difference according to the rank**
I	II	III	II	III	II	III	II	III
1	GAENE 5	0.53	GAENE 7	0.03	GAENE 7	0.21	GAENE 5	0. 38
2	GAENE 3	0.25	GAENE 3	0.19	GAENE 5	0.34	GAENE 13	0.12
3	GAENE 4	0.18	GAENE 5	0.24	GAENE 13	0.10	GAENE 2	0.10
4	GAENE 7	0.13	GAENE 4	0.03	GAENE 4	0.17	GAENE 1	0.39
5	GAENE 1	0.01	GAENE 1	0.14	GAENE 3	0.05	GAENE 11	0.07
6	GAENE 13	0.30	GAENE 13	0.07	GAENE 1	0.03	GAENE 12	0.25
7	GAENE 2	0.11	GAENE 2	0.18	GAENE 9	0.26	GAENE 4	0.08
8	GAENE 9	0.32	GAENE 9	0.21	GAENE 2	0.12	GAENE 3	0.10
9	GAENE 11	0.36	GAENE 11	0.16	GAENE 10	0.14	GAENE 9	0.01
10	GAENE 10	0.35	GAENE 10	0.19	GAENE 11	0.71	GAENE 10	0.09
11	GAENE 12	0.49	GAENE 12	0.13	GAENE 6	0.09	GAENE 7	0.34
12	GAENE 8	0.03	GAENE 6	0.08	GAENE 12	0.14	GAENE 6	0.40
13	GAENE 6		GAENE 8		GAENE 8		GAENE 8	

Note. The differences ranking may, according to the Rasch model, be an indication of the construct validity of the instrument

Each category is more likely (dotted line in **Figure 3**) to appear in a particular combination of difference between person measures and item difficulty (Boone et al., 2014; Smith et al., 2002), with no disordering in the sequence of the categories. For example, at -2.00 logits, the probability of observing the "disagree" category is about 60%, while the probability of observing the categories "strongly disagree" or "I am uncertain/I cannot decide" is less than 20%. In addition to, at +2.00 logits, the probability of observing the "agree" category is approximately 70%, while the probability of observing the "strongly agree" category is less than 20% (Smith et al., 2002). **Figure 3** (dushed line) also shows Andrich thresholds for each pair of categories, which have sufficient spacing (SD-D=-2.98, D-U=-0.98, U-A=+0.59, A-SA=+3.37) and regular order between them (SD-D<D-U<U-A<A-SA), indicating the satisfactory operation of the scale.

The index person reliability (**Table 5**) was 0.77, which in combination with the separation index=1.85 means that it is not easy to separate person in terms of the levels of acceptance of evolution. This sample can separate person into a maximum of three levels of acceptance (Linacre, 2021). On the contrary, the index item reliability was 0.94, which means that the sample size is large enough to confirm the hierarchy of the tool's items in terms of the degree of difficulty or convenience of the respondents in terms of its items. Also, the separation index was 4.02 indicating that the tool can distinguish four levels of difficulty between the items of the instrument for this sample (Smith et al., 2002).

The hierarchy of the items of the instrument in the sample is different than the hierarchy in the surveys that were conducted in the past with the GAENE 2.1 (**Table 6**). **Table 6** shows data concerning the order of GAENE 2.1 items according to the ranking from the easiest to the most difficult to agree with (column I: MEASURE ORDER). The ranking of each item in the different studies are presented in column II, which shows that some items have different ranking in the different studies. Column III of each study are presented the differences between two consecutive rankings. These differences reveal that the intervals between the items are small and unequal.

Typical examples for the hierarchy of the items are GAENE 3, GAENE 7, and GAENE 13. In terms of the item GAENE 3, there is a different classification in the present study from that of both Sbeglia and Nehm (2018) and the research of Smith et al. (2016) and of Romine et al. (2017). The item GAENE 7 has the same ranking in the research of Romine et al. (2017) and Sbeglia and Nehm (2018) and it is the one that is most easily agreed on, while in the research of Smith et al. (2016) this item is about in the middle in terms of convenience, which one can agree with. In the present research it seems to be one of the most difficult to agree on. In the

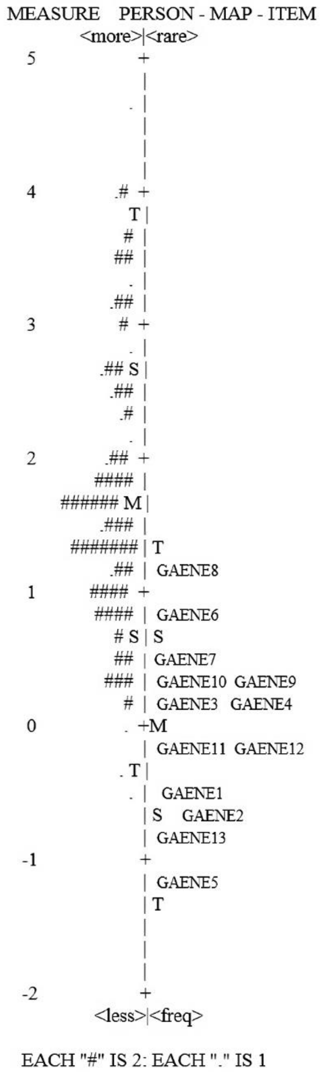

Figure 4. Wright map (map expresses in the same unit of measurement [logit], both ability of respondent to answer [left part of map], and degree of difficulty of item [right part of map]; persons marked with an X, & those with the lowest ability are at the bottom left of map, while those with the highest ability are at the top left of the map; the item that one can easily agree with is at the bottom right of the map [GAENE5], while the one which one can hardly agree with is at the top right of the map [GAENE6]) (Winsteps: Rasch Measurement Computer Program)

item GAENE 13 the present study agrees with that of Sbeglia and Nehm (2018) but not with the research of Smith et al. (2016) and Romine et al. (2017).

In this research, the item with which one can easily agree with is GAENE 5, while in the item with which one can hardly agree with is GAENE 8 (**Figure 4**: Wright map). The map shows that the items of the instrument have a dense distribution and in some there are overlaps. The range of difficulty of the data is relatively limited as it ranges from -1.16 to +1.17. The range of difficulty of two consecutive items of the instrument ranges from 0.01 to 0.40 indicating the age homogeneity of the sample (Boone et al., 2014; DeMars & Linacre, 2004).

RQ2

Table 7 presents the results of students' acceptance of evolution. The MEASURE column shows the levels of acceptance of evolution by each respondent in descending order according to Rasch model. TOTAL SCORE column shows the level of acceptance of evolution of each respondent using the raw data. The average value of students' responses to the raw data is MEAN=49.10 (MEAN by ITEM=3.82/5). In previous studies, with the GAENE, the levels of the acceptance of evolution have a range in raw data from 47.06 to 55.38 (Metzger et al., 2018; Sbeglia & Nehm, 2018; Smith et al., 2016; Rachmatullah et al., 2018). The raw data score of 49.10, is the second lowest score among the studies that have been conducted with GAENE. The latter can be justified by the fact that its application was made to students who have little knowledge of the living world in general and of evolution in particular. However, the results may not be directly comparable between these surveys because each of them was applied in different populations.

The levels of the acceptance of evolution in the present study do not appear to be influenced by gender (ANOVA: $f_{1,115}$=3.58, p=0.058) and the field of study they intend to pursue in the next class (ANOVA: $f_{1,113}$=0.48, p=0.69). On the contrary, they seem to be influenced by whether they have attended evolution courses in the third grade of junior high school.

Table 7. Person measures

EN	TS	M	MSE	Infit		Outfit		EN	TS	M	MSE	Infit		Outfit	
				MNSQ	ZSTD	MNSQ	ZSTD					MNSQ	ZSTD	MNSQ	ZSTD
86	62	4.73	.66	1.61	1.31	1.29	.67	77	49	1.46	.43	.64	-.85	.64	-.86
66	60	4.00	.56	.76	-.64	.71	-.74	82	49	1.46	.43	.25	-2.57	.23	-2.64
78	60	4.00	.56	1.33	.97	1.18	.57	96	49	1.46	.43	.61	-.98	.54	-1.20
99	60	4.00	.56	1.44	1.20	1.37	1.00	102	49	1.46	.43	1.03	.21	1.10	.38
34	59	3.71	.53	.86	-.30	.90	-.18	112	48	1.28	.42	.36	-1.97	.35	-2.02
35	59	3.71	.53	.93	-.10	.95	-.03	1	48	1.28	.42	1.24	.68	1.22	.64
18	58	3.44	.51	1.06	.29	1.03	.21	10	48	1.28	.42	1.14	.46	1.07	.30
39	58	3.44	.51	1.20	.65	1.16	.56	15	48	1.28	.42	.52	-1.30	.49	-1.43
50	58	3.44	.51	1.17	.57	1.16	.56	16	48	1.28	.42	1.07	.31	.95	.01
91	58	3.44	.51	1.29	.87	1.20	.67	22	48	1.28	.42	.64	-.87	.61	-.99
88	53	3.26	.52	1.29	.85	1.30	.88	30	48	1.28	.42	1.34	.89	1.36	.94
7	57	3.18	.50	.62	-1.10	.64	-1.06	33	48	1.28	.42	.81	-.38	.79	-.41
44	57	3.18	.50	1.07	.30	1.06	.27	65	48	1.28	.42	.71	-.67	.71	-.66
56	57	3.18	.50	.48	-1.70	.48	-1.72	80	48	1.28	.42	1.41	1.03	1.26	.73
114	57	3.18	.50	.90	-.17	.87	-.27	87	48	1.28	.42	1.00	.13	1.03	.21
38	52	3.11	.51	1.86	1.95	1.79	1.87	89	48	1.28	.42	.96	.03	.92	-.08
8	56	2.94	.49	1.44	1.16	1.38	1.05	101	48	1.28	.42	.58	-1.08	.54	-1.23
52	56	2.94	.49	1.43	1.15	1.40	1.10	108	48	1.28	.42	2.05	2.13	2.09	2.18
73	51	2.78	.50	1.84	1.83	1.72	1.66	110	48	1.28	.42	.84	-.27	.78	-.44
4	55	2.71	.48	1.48	1.21	1.45	1.17	12	47	1.10	.42	.73	-.60	.71	-.67
57	55	2.71	.48	.75	-.59	.75	-.59	27	47	1.10	.42	.84	-.28	.87	-.19
79	55	2.71	.48	.51	-1.45	.54	-1.36	31	47	1.10	.42	.62	-.98	.60	-1.02
83	55	2.71	.48	.92	-.09	.91	-.13	54	47	1.10	.42	2.07	2.19	2.15	2.30
28	50	2.61	.49	1.98	2.00	1.98	2.02	74	47	1.10	.42	.75	-.55	.73	-.62
84	46	2.55	.51	1.17	.51	1.13	.44	47	36	1.04	.47	.59	-.92	.51	-1.15
51	42	2.55	.54	1.88	1.73	1.75	1.57	14	46	.93	.41	.48	-1.54	.51	-1.38
63	54	2.48	.47	1.05	.27	1.16	.51	25	46	.93	.41	.44	-1.68	.48	-1.49
92	54	2.48	.47	.88	-.19	.89	-.17	49	46	.93	.41	1.72	1.64	1.72	1.63
100	54	2.48	.47	2.27	2.49	2.25	2.51	71	46	.93	.41	.78	-.47	.77	-.48
3	53	2.26	.46	.77	-.47	.71	-.66	95	46	.93	.41	.87	-.22	.90	-.14
5	53	2.26	.46	1.38	.96	1.36	.94	98	46	.93	.41	.62	-.98	.63	-.95
105	53	2.26	.46	1.28	.76	1.20	.60	104	46	.93	.41	.48	-1.51	.44	-1.66
60	48	2.14	.48	.41	-1.61	.40	-1.68	11	42	.82	.42	.58	-1.09	.60	-.99
115	48	2.07	.47	.85	-.24	.87	-.19	32	35	.81	.46	.84	-.23	.80	-.33
2	52	2.05	.46	.94	-.02	.86	-.23	20	45	.76	.40	.81	-.40	.77	-.52
19	52	2.05	.46	.92	-.07	.92	-.08	23	45	.76	.40	1.93	2.04	1.88	1.93
29	52	2.05	.46	1.48	1.14	1.47	1.13	40	45	.76	.40	1.03	.21	1.01	.14
53	52	2.05	.46	1.69	1.53	1.68	1.52	48	45	.76	.40	.53	-1.36	.49	-1.50
24	51	1.85	.45	1.00	.15	1.00	.13	58	45	.76	.40	.31	-2.37	.31	-2.37
62	51	1.85	.45	.14	-3.29	.14	-3.38	107	45	.76	.40	1.14	.47	1.13	.46
72	51	1.85	.45	.65	-.81	.65	-.84	36	44	.60	.40	1.17	.55	1.16	.53
81	51	1.85	.45	.41	-1.72	.42	-1.66	67	44	.60	.40	.84	-.31	.86	-.24
85	51	1.85	.45	.97	.07	.97	.06	43	43	.45	.39	.34	-2.31	.34	-2.25
109	51	1.85	.45	2.17	2.25	2.16	2.25	45	43	.45	.39	.45	-1.74	.45	-1.74
117	51	1.85	.45	.91	-.08	.93	-.02	55	43	.45	.39	.77	-.54	.74	-.64
13	47	1.82	.47	1.00	1.80	1.90	1.80	75	43	.45	.39	.1.76	1.78	1.84	1.91
6	50	1.65	.44	1.35	.89	1.33	.86	17	42	.29	.39	.87	-.25	.83	-.35
21	50	1.65	.44	1.20	.60	1.09	.34	41	42	.29	.39	.94	-.05	.90	-.17
37	50	1.65	.44	1.08	.32	1.11	.39	68	42	.29	.39	.47	-1.69	.47	-1.68
42	50	1.65	.44	.91	-.09	.98	.09	76	42	.29	.39	1.30	.87	1.40	1.08
64	50	1.65	.44	.55	-1.16	.56	-1.15	94	42	.29	.39	2.10	2.38	2.04	2.27
90	50	1.65	.44	.39	-1.79	.36	-1.93	97	42	.29	.39	1.30	.87	1.40	1.08
93	50	1.65	.44	1.07	.30	1.02	.18	26	41	.14	.39	1.38	1.04	1.37	1.01
103	50	1.65	.44	1.09	.35	1.00	.15	70	41	.14	.39	.58	-1.23	.58	-1.24
106	50	1.65	.44	.89	-.13	.92	-.06	61	40	-.01	.38	1.03	.21	1.09	.36
111	50	1.65	.44	.26	-2.47	.25	-2.55	46	38	-.30	.38	1.10	.40	1.09	.37
113	50	1.65	.44	1.02	.19	1.02	.18	59	37	-.44	.38	1.44	1.19	1.49	1.29
116	50	1.65	.44	.97	.07	.97	.06	MEAN	49.10	1.68	.45	1.02	-.01	1.00	-.05
9	49	1.46	.43	.54	-1.20	.58	-1.07	P.SD	5.4	1.04	.05	.47	1.19	.46	1.19
69	49	1.46	.43	1.01	.17	1.06	.29								

Note. EN: Entry number; TS: Total score; M: Measure; MSE: Model SE;

DISCUSSION AND CONCLUSIONS

GAENE 2.1 is a psychometric tool that measures a theoretical construction, the acceptance of evolution, in different populations. Its items measure a person's willingness to support evolution (Barnes et al., 2019). According to the researchers who developed it, it does not contain items that are related to the understanding of evolution (Smith et al., 2016). Previous research that conducted with undergraduate students shown that the GAENE raw scores increased from pre- to the post-courses (from 52.78/65 to 55.38/65) but did not reveal statistically significant difference in the levels of the acceptance of evolution (Sbeglia & Nehm, 2018). In this study seems to be a statistically significant difference between students who have taken evolutionary courses compared to those who have not taken evolutionary courses (ANOVA: f1,115=3.90, p=0.048) in junior high school.

Further analysis of the data showed differential item functioning (DIF)for the GAENE 4 (DIF contrast: -1.19, prob.= 0.00), GAENE 5 (DIF contrast: -0.68, prob.=0.05) and GAENE 8 (DIF contrast: 0.66, prob.=0.01) between the two subgroups. Students who had attended evolutionary courses in the third grade of junior high school were more likely to agree with GAENE 4 and GAENE 5 than students who had not attended the respective courses, while the reverse picture was observed for GAENE 8.

Previous research conducted on a sample of college students showed that item GAENE 12 had pre-course fit indicator MNSQ out of range (infit: 1.46; outfit: 1.51) and it had post-course fit indicator MNSQ out of range too (infit: 2.24; outfit: 2.10), suggesting that this item may be influenced by levels of knowledge of the evolutionary theory (Sbeglia & Nehm, 2018). This item had a problematic behavior during the initial implementation of the GAENE 2.1 as it showed fit MNSQ indicators out of limits (Infit: 1.43; Outfit: 1.42) and a significant difference (1.00 logits in DIF) among students who had the same levels of acceptance of the evolution, which it means that the item should be affected by the education level (Smith et al., 2016).

A possible explanation for the above findings lies in the wording of the item, as for evolutionary biologists evolution is both a biological phenomenon and a scientific theory too (Sbeglia & Nehm, 2018). In this research, GAENE 12 has fit indicator MNSQ within limits (infit:1.06; outfit: 1.07) and there does not seem to be a significant difference between students who have taken compared to those who have not taken evolutionary courses (DIF contrast: 0.00, prob.=1.00) in junior high school. The behavior of the sample in this item can be explained:

(a) by the fact that all participants of the sample are in the same educational level,

(b) by the lack of substantial knowledge about the evolution of the students of this age group in Greece despite the fact that some aspects of the evolution may have been taught such as natural selection and adaptation (Prinou et al., 2008), and

(c) by the lack of knowledge about the nature of science in Greek students of that age (Koumara & Plakitsi, 2020).

DIF analysis by gender was shown that all items had not significant difference between two subgroups (DIF contrast: -0.56-+0.39, prob.=0.08-0.81) with the exception of the GAENE 11 (DIF contrast: -0.85, prob.=0.01). A possible explanation is the girls in the aged of 17 are more sensitive in human evolution than the boys.

The data drawn from the instruments and especially those that were submitted to content analysis paved the way to further the research on students of the specific age group before and after having been exposed to teaching of evolution.

The analysis of the quality of GAENE as a tool for measuring the acceptance of evolution during the application in the specific sample includes the evaluation of the behavior of the categories of the scale.

Table 8 shows the classifications of each category of the scale per item and shows their normal distribution and the existence of a sufficient distance between them, with the sole exception of GAENE 11. In this item there is an apparent disturbance in the order of categories due to the small number of people who have chosen category 2 "disagree" in relation to people who have chosen category 3 "I am uncertain/I cannot decide". Thus, it can be concluded from the results that for the specific sample, the categories of the scale work with the expected predictability (Boone et al., 2014) (**Table 8**).

From the Wright map (**Figure 4**) it is understood that the instrument consists of items that are easy to agree with, because the average of the students' answers is much higher than the average of the difficulty of the items (indication M left and right of the dashed line in **Figure 4**, respectively) (Boone, 2016). The same picture is observed in the initial research of the GAENE 2.1 (Smith et al., 2016 p. 19) and in another research too (Sbeglia & Nehm, 2018 p. 13/19). This finding may mean that the GAENE 2.1 overestimates students' acceptance of evolution and may explain the difference between the public and the students' acceptance of evolution.

In addition, in all published studies that conducted with the GAENE 2.1, the items, which one can hardly agree with are the GAENE 8, GAENE 6, and GAENE 10 (**Table 3**). However, items GAENE 6 and GAENE 10 have a similar wording, and the degree of agreement may not depend on the acceptance of evolution but on other factors related to the personality of the respondent and his or her desire to argue publicly about an issue (Sbeglia & Nehm, 2018) or the fear one may feel when speaking in public (Romine et al., 2018). GAENE 8 is essentially the famous phrase of Dobzhansky (1973), it includes the word "nothing" that could be treated by students with reservation. This may be due to the fact that teachers often and especially for biological phenomena instruct students to avoid words like everything, always, nothing etc. (Sbeglia & Nehm, 2018).

Also, the different hierarchy in the difficulty of the items observed in the different surveys conducted with GAENE (**Table 6**), the dense distribution of items and the existence of overlaps between them (**Figure 3**), indicate that in some of these items they need to be reworded in order to increase the range of distribution according to their difficulty and at the same time to have a more normal distance between them (Arnold et al., 2018). Attributing a dense distribution in an instrument that measures a variable may mean that the full range of the variable is not covered by the items, while the existence of overlaps can affect the tool's ability to measure the variable in an efficient and effective way (Boone et al., 2014).

Table 8. Item category frequencies (rating scale category frequencies)

Item	DCode	DC	Ability			Item	DCode	DC	Ability		
			Mean	PSD	SE mean				Mean	PSD	SE mean
GAENE 1	1					GAENE 8	1	1	-0.44	0.00	
	2	3	-0.25	0.18	0.13		2	23	1.06	0.74	0.16
	3	12	1.04	0.53	0.16		3	36	1.51	0.85	0.14
	4	78	1.56	0.86	0.10		4	50	2.02	1.06	0.15
	5	24	2.65	1.02	0.21		5	6	2.81	0.53	0.24
GAENE 2	1					GAENE 9	1				
	2	1	-0.44	0.00			2	8	0.60	0.68	0.26
	3	14	1.05	0.61	0.17		3	32	1.39	0.80	0.14
	4	76	1.56	0.92	0.11		4	60	1.69	0.97	0.13
	5	26	2.49	1.03	0.21		5	13	2.84	0.93	0.27
GAENE 3	1					GAENE 10	1				
	2						2	11	0.50	0.43	0.14
	3	44	1.19	0.76	0.12		3	24	1.27	0.81	0.17
	4	57	1.73	0.94	0.13		4	70	1.78	0.85	0.10
	5	14	2.97	1.05	0.29		5	11	3.24	0.94	0.30
GAENE 3	1					GAENE 11	1	1	0.29	0.00	
	2	4	0.25	0.07	0.04		2	1	1.10	0.00	
	3	41	1.23	0.76	0.12		3	30	1.01*	0.64	0.12
	4	47	1.54	0.70	0.10		4	58	1.56	0.84	0.11
	5	23	3.03	0.90	0.19		5	24	2.90	0.89	0.19
GAENE 4	1					GAENE 12	1				
	2	3	0.34	0.07	0.05		2	4	0.68	0.67	0.39
	3	11	0.98	0.62	0.20		3	32	1.01	0.66	0.12
	4	55	1.29	0.66	0.09		4	53	1.68	0.90	0.12
	5	45	2.37	1.10	0.17		5	26	2.68	0.89	0.18
GAENE 6	1	2	0.45	0.31	0.31	GAENE 13	1				
	2	17	0.72	0.51	0.13		2				
	3	35	1.27	0.75	0.13		3	16	0.79	0.68	0.17
	4	46	1.98	0.82	0.12		4	69	1.46	0.80	0.10
	5	17	2.86	1.05	0.26		5	31	2.67	0.93	0.17
GAENE 7	1										
	2	3	1.13	0.63	0.45						
	3	45	1.39	1.19	0.18						
	4	58	1.80	0.85	0.11						
	5	11	2.40	0.83	0.26						

Note. DCode: Data code & DC: Data count

Table 9. GAENE 2.1. quality criteria

Criterion	Value	Designation (Fisher, 2007)	Cut-off for excellent performance (Fisher, 2007)
Variance in data explained by measures	35.2%	Poor	>80%
Person reliability (Cronbach's alpha)	0.77	Fair	>0.94
Person separation index	1.85	Poor	>5.00
Item reliability	0.94	Very good	>0.94
Item Separation index	4.02	Very good	>5.00
Item infit (MNSQ)	0.72-1.25	Very good–excellent	0.77-1.30
Item outfit (MNSQ)	0.74-1.27	Very good–excellent	0.77-1.30
Ceiling effect	0%	Excellent	<3%
Floor effect	0%	Excellent	<3%

Finally, the evaluation of the implementation of GAENE as a tool for measuring the acceptance of evolution in Greek students aged 17 years, can be done with various criteria (**Table 9**) (Fisher, 2007). One such criterion is the percentage of variance in the data explained by the measurements of individuals and data (**Figure 1**), which for this sample is low (35.8%) and may be due to unexpected student responses (**Table 7** see INFIT and OUTFIT columns) in specific items of the tool. It may, also, be due to the lack of essential knowledge about the biological world and the evolution of organisms (Prinou et al., 2008). The person reliability is relatively low (0.77) that in combination with the low separation index (1.85) indicates that GAENE can distinguish three levels of acceptance among the respondents (Linacre, 2021). In contrast, the Item reliability and the corresponding separation index are relatively high (0.94 and 4.02 respectively), which means that the sample is large enough to confirm the hierarchy of item difficulty (**Table 6**) and that the items of the tool can be distinguished in at least four levels of difficulty (Smith et al., 2002). The fit indicators are within the strict limits of the excellent fit,–with the exception of only two items–proving the very good fit of the tool to Rasch model. Also, there are no extreme values, i.e., there are no people who have achieved the lowest and highest possible result.

The studies conducted in Greece on the acceptance of evolution, have exclusively used MATE as measurement tool and the range of the results in these studies was from 70.95 to 90.5. These studies were conducted on students in pedagogical and biological departments of Greek universities, but also on teachers at primary schools, junior and senior high schools (Athanasiou et al., 2012; Athanasiou & Papadopoulou, 2012; Katakos & Athanasiou, 2020; Mantelas & Mavrikaki, 2020). Thus, the results of the current research are not directly comparable with these, not only because a different measurement tool was used, but also because it was applied to 17 years old students. However, if the results are equalized with MATE scores then this study obtains 75.54 score. This score is higher than the minimum resulting from the already mentioned surveys.

The levels of acceptance of evolution balanced with MATE, in this research, do not show a significant difference with corresponding research that have been conducted in students at Greek universities (Athanasiou et al., 2012, 2016; Athanasiou & Papadopoulou, 2012; Athanasiou & Papadopoulou, 2015; Mantelas & Mavrikaki, 2020), because GAENE does not contain items related to the knowledge of evolutionary theory and religiosity (Smith et al., 2016). Recent research conducted on Biology students in Greece has shown that there is a negative correlation between religiosity and levels of acceptance and that attending evolutionary courses is one of the factors that has a significant positive impact on acceptance of evolution, at least when MATE is used as a measurement instrument (Mantelas & Mavrikaki, 2020).

At the same time, opinion polls conducted on a sample of young people aged 15-29 in Greece showed that young people have a relaxed relationship with religious practices (Special Eurobarometer 225, 2014), which may partly explain the difference observed in the levels of acceptance of evolution in adults (Miller et al., 2006). Another parameter that may be related to the difference in the levels of acceptance of evolution between young people and adults is the formulation of the proposals (Smith et al., 2016). In the opinion polls undertaken on this issue, the formulations of the proposals also concern the intervention of God in the evolutionary process and the appearance of man (Gallup, 2019; Miller et al., 2006).

Thus, taking into account the above findings of the research and the peculiarities of the Greek educational system regarding the teaching of evolution (Prinou et al., 2008; Stasinakis & Kampourakis, 2018), we can conclude that GAENE, despite the fact that it is likely to overestimate the levels of the acceptance of evolution, it is a reliable tool to measure this variable in the specific age group, because it does not contain items related to the knowledge and understanding of evolutionary theory. In addition, the present study proposes the further monitoring of the items of the tool during its application in a larger sample, which covers a larger geographical area and includes students from wider social backgrounds.

Limitations of Research

The present research was undertaken in a small population that has not been taught evolution neither as a phenomenon of the living world, nor as a scientific theory. Also, the structure of the educational system in Greece is such that essentially it does not allow the development of biological knowledge in students, because from an early age, pupils are taught how to succeed in the entrance exams for university and are indifferent to courses that deviate from this target (Stasinakis & Kampourakis, 2018). An additional important limitation for the present study is the restricted area in which it took place, and this might have some effect on the results as discussed in a previous section. Thus, its results are not generalizable beyond the population studied but they yield some interesting findings that can be a guide for further research, in a wider area with a larger sample, which includes students from different regions, and different social and cultural backgrounds. Such a research should evaluate each item of GAENE separately, in order to evaluate GAENE more accurately in terms of the ability of this tool to effectively measure acceptance of evolution in Greek students at high school. It is also important to evaluate the implementation of the various tools for measuring the acceptance of evolution to Greek students in order to make comparative studies between different populations and different tools (GAENE 2.1, MATE, and I-SEA), which will help to improve these tools and our ability to measure the acceptance of evolution.

Author contributions: All authors have sufficiently contributed to the study and agreed with the results and conclusions.

Funding: No funding source is reported for this study.

Ethical statement: The authors stated that, the principal of the educational units and the teachers of the classes were aware of the content of the study and the questionnaires, the questionnaires were anonymous and the participation of the respondents was completely voluntary.

Declaration of interest: No conflict of interest is declared by authors.

Data sharing statement: Data supporting the findings and conclusions are available upon request from the corresponding author.

REFERENCES

AAAS. (2011). *Vision and change in undergraduate biology education: A call to action*. AAAS.

AERA. (2014). *Standards for educational and psychological testing*. AERA

Andrich, D., & Marais, I. (2019). *A course in Rasch Measurement Theory: Measurement in the educational, social and health sciences*. Springer Nature Singapore Pte Ltd. https://doi.org/10.1007/978-981-13-7496-8

Arnold, J. C., Boone W. J., Kremer K., & Mayer J. (2018). Assessment of competencies in scientific inquiry through the application of Rasch measurement techniques. *Education and Sciences, 8*(184), 1-20. https://doi.org/10.3390/educsci8040184

Athanasiou, K., & Papadopoulou, P. (2012). Conceptual ecology of the evolution acceptance among Greek education students: Knowledge, religious practices and social influences. *International Journal of Science Education, 34*(6), 903-924. https://doi.org/10.1080/09500693.2011.586072

Athanasiou, K., & Papadopoulou, P. (2015). Evolution theory teaching and learning: What conclusions can we get from comparisons of teachers; and students' conceptual ecologies in Greece and Turkey? *EURASIA Journal of Mathematics, Science & Technology Education, 11*(4), 841-853. https://doi.org/10.12973/Eurasia.2015.1443a

Athanasiou, K., Katakos, E., & Papadopoulou, P. (2012). Conceptual ecology of the evolution acceptance among Greek education students: The contribution of knowledge increase. *Journal of Biological Education, 46*(4), 234-241. https://doi.org/10.1080/00219266.2012.716780

Athanasiou, K., Katakos, E., & Papadopoulou, P. (2016). Acceptance of evolution as one of the factors structuring the conceptual ecology of the evolutionary theory of Greek secondary school teachers. *Evolution: Education and Outreach, 9*(1), 7. https://doi.org/10.1186/s12052-016-0058-7

Barnes, E. M., Dunlop, H. M., Holt, E. A., Zheng, Y., & Brownell, S. E. (2019). Different evolution acceptance instruments lead to different research findings. *Evolution: Education and Outreach, 12*(4), 1-17. https://doi.org/10.1186/s12052-019-0096-z

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? *Life Sciences Education, 15*(4), 1-7. https://doi.org/10.1187/cbe.16-04-0148

Boone, W. J., & Noltemeyer, A. (2017). Rach analysis: A primer for school psychology researchers and practitioners. *Cogent Education, 4*(1416898), 1-13. https://doi.org/10.1080/2331186X.2017.1416898

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer. https://doi.org/10.1007/978-94-007-6857-4

Creswell, J. W. (2012). *Educational research: Planning, conducting and evaluating quantitative and qualitative research* (4th Edn.). Pearson Education Inc.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334. https://doi.org/10.1007/bf02310555

Dawkins, R. (2009). *The greatest show on earth*. FreePress, TransWorld.

DeMars, C., & Linacre, J. M. (2004). Mapping multidimensionality. *Rasch Measurement Transactions, 18*(3), 9990-9991.

Dobzhansky, T. (1973). Nothing in biology makes sense except in the light of evolution. *The American Biology Teacher, 35*(3), 125-129. https://doi.org/10.2307/4444260

Dunk, R. D. P., Barnes, M. P., Reiss, M. J., Alters, B. Asghar, A., Carter, B. E., Cotner, S., Glaze, A. L., Hawley, P. H., Jensen, J. L., Mead, L. S., Nadelson, L. S., Nelson, C. E., Pobiner, B., Scott, E. C., Shtulman, A., Sinatra, G. M., Southerland, S. A., Walter, E. M., … Dunk, J. R. W. (2019). Evolution education is a complex landscape. *Nature, Ecology and Evolution, 3*, 327-329. https://doi.org/10.1038/s41559-019-0802-9

Eder, E., Seidl, V., Lange, J., & Dittmar, G. (2018). Evolution education in the German-speaking countries. In D. Hasan, & L. A. Borgerding (Eds.), *Evolution education around the globe* (pp. 235-260). Spinger. https://doi.org/10.1007/978-3-319-90939-4_13

Fisher, J. R. (2007). Rasch measurement transaction. *Transaction of Rasch Measurement, 21*(1), 1095.

Gallup. (2019). Evolution, creationism, intelleigent design. *Gallup.com*. https://news.gallup.com/poll/21814/evolution-creationism-intelligent-design.aspx?version=print

Gardner, P. L. (1995). Measuring attitubes to science: Uni-dimensionality and internal consistency revised. *Research in Science Education, 25*(3), 283-289. https://doi.org/10.1007/bf02357402

Gregory, T. R. (2009). Understanding natural selection: Essential concepts and common misconceptions. *Evolution: Education and Outreach, 2*(2), 156. https://doi.org/10.1007/s12052-009-0128-1

Hagell, P. (2014). Testing rating scale unidimensionality using the Principal Component Analysis (PCA)/t-test protocol with the Rasch model: The primacy of theory over statistics. *Open Journal of Statistics, 4*, 456-465. https://doi.org/10.4236.ojs.2014.46044

Harms, U., & Reiss, M. J. (2019). The present status of evolution education. In U. harms, & M. J. reiss (Eds.), *Evolution education re-considered: Understanding what works* (pp. 1-20). Springer. https://doi.org/10.1007/978-3-030-14698-6

Heale, R., & Twycross, A., (2015). Validity and reliability in quantitative studies. *Evidence-Based Nursing, 18*(3), 66-67. https://doi.org/10.1136/eb-2015-102129

Kampourakis, K., & Zogza, V. (2008). Students' intuitive explanations of the causes of homologies and adaptations. *Science & Education, 17*(1), 27-47. https://doi.org/10.1007/s11191-007-9075-9

Katakos, S., & Athanasiou, K. (2020). The 'geological argument' as an Instrument for the acceptance of the theory of evolution among Greece science teachers. *Journal of Cell Biology and Genetics, 3*(3), 183-186.

Kazempour, M., & Amirshokoohi, A., (2018). Evolution education in Iran: Shattering myths about teaching evolution in an Islamic state. In D. Hasan, & L. A. Borgerding (Eds.), *Evolution education around the globe* (pp. 281-295). Spinger. https://doi.org/10.1007/978-3-319-90939-4_15

Knapp, T. R. (1990). Treating ordinal scales as interval scales an attempt to resolve the controversy. *Nursing Research, 39*(2), 121-123. https://doi.org/10.1097/00006199-199003000-00019

Koumara, A., & Plakitsi, K., (2020). The degree that nature of scientific knowledge aspects are included in the science class of Greek high schools. *World Journal of Education, 10*(5), 1-17. https://doi.org/10.5430/wje.v10n5p1

Kuschmierz, P., Beniermann, A., Bergmann, A., Pinxten, R., Aivelo, T., Berniak-Wozny, J., Bohlin, G., Bugallo-Rodriguez, A., Cardia, P., Cavadas, B. F. B. P., Cebesoy, U. B., Cvetkovic, D. D., Demarsy, E., Dordevic, M. S., Drobniak, S. M., Dubchak, L., Dvorakova, R. M., Fancovicova, J., Fortin, … Graf, D. (2021). European first-year university students accept evolution but lack substantial knowledge about it: A standardized European cross-country assessment. *Evolution: Education and Outreach, 14*, 17. https://doi.org/10.1186/s12052-021-00158-8

Kuzon, W. M., Urbanchek, M. G., & McCabe, S. (1996). The seven deadly sins of statistical analysis. *Annals of Plastic Surgery, 37*(3), 265-272. https://doi.org/10.1097/00000637-199609000-00006

Lennox, G. J. (1992). The philosophy of biology. In S. H. Merrilee (Ed.), *Introduction to the philosophy of science.* Prentice-Hall.

Linacre, J. M. (1997). KR-20 or Rasch reliability: Which tells the "truth"? *Rasch Measurement Transactions, 11*(3), 580-581.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2022). WINSTEPS. Rasch measurement computer program. *Winsteps.com*. https://www.winsteps.com/winman/copyright.htm

Linacre, J. M., (2021). A user's guide to Winsteps ministeps Rasch-model computer programs. *Winsteps.com*. http://www.winsteps.com/winsteps.htm

Mantelas, N., & Mavrikaki, E. (2020). Religiosity and students' acceptance of evolution. *International Journal of Science Education, 42*(18), 3071-3092. https://doi.org/10.1080/09500693.2020.1851066

Mead, L. S., Kohn, C., Warwick, A., & Schwartz, K., (2019). Applying measurement standards to evolution education assessment instruments. *Evolution: Education & Outreach, 12*(5), 1-14. https://doi.org/10.1186/s12052-019-0097-y

Metzger, K. J., Montplairir, D., Haines, D., & Nickodem, K. (2018). Investigating undergraduate health sciences students' acceptance of evolution using MATE and GAENE. *Evolution: Education & Outreach, 11*(10), 1-18. https://doi.org/10.1186/s12052-018-0084-8

Miller, J. D., Scott, E. C., & Okamoto S. (2006). Public acceptance of evolution. *Science, 313*(5788), 765-766. https://doi.org/10.1126/science.1126746

NGSS Lead States. (2013). *Next generation science standards: For states by states.* The National Academies Press.

Planinic, M., Boone, W. J., Susac, A, & Ivanjek, L. (2019). Rasch analysis in physics education research: Why measurement matters. *Physical Review Physics Education Research, 15*(020111), 1-14. https://doi.org/10.1103/PhysRevPhysEducRes.15.020111

Prinou, L., Halkia, K., & Skordoulis, C. (2008). What conceptions do Greek school students form about biological evolution? *Evolution: Education & Outreach, 1*(3), 312-317. https://doi.org/10.1007/s12052-008-0051-x

Quessada, M.-P., & Clément, P. (2018). Evolution education in France: Evolution is widely taught and accepted. In D. Hasan, & L. A. Borgerding (Eds.), *Evolution education around the globe* (pp. 213-233). Spinger. https://doi.org/10.1007/978-3-319-90939-4_12

Rachmatullah, A., Nehm, R., Roshayanti, F., & Ha, M. (2018). Evolution education in Indonesia: Pre-service biology teacher's knowledge, reasoningh models and acceptance of evolution. In D. Hasan, & L. A. Borgerding (Eds.), *Evolution education around the globe* (pp. 335-355). Spinger. https://doi.org/10.1007/978-3-319-90939-4_18

Raiche, G. (2005). Critical eigevalue sizes (variances) in standardized residual principal components analysis. *Rasch Measurement Transactions, 19*(1), 1012.

Romine, W. L., Todd, A. N., & Walter, E. M. (2018). A closer look at the items within three measures of evolution acceptance: Analysis of the MATE, I-SEA, and GAENE aw a single corpus of items. *Evolution: Education and Outreach, 11*, 17. https://doi.org/10.1186/s12052-018-0093-7

Romine, W. L., Walter, E. M., Bosse E., & Todd A. N. (2017). Understanding patterns of evolution acceptance-a new implementantion of the measure of acceptance of the theory of evolution (MATE) with Midwestern University students. *Journal of Research in Science Teaching, 54*(4), 642-671. https://doi.org/10.1002/tea.21380

Rutledge, M. L., & Warden, M. A. (1999). The development and validition of the measure of acceptance of the theory of evolution instrument. *School Science and Mathematics Journal, 99*(1), 13-18. https://doi.org/10.1111/j.1949-8594.1999.tb17441.x

Rutledge, M., & L., Sadler, K. C. (2007). Reliability of the measure of acceptance of the theory of evolution (MATE) instrument with university students. *The American Biology Teacher, 69*(6), 332-335. https://doi.org/10.1662/0002-7685(2007)69[332:ROTMOA]2.0.CO;2

Sbeglia, G. C., & Nehm, R. H. (2018). Measuring evolution acceptance using the GAENE: Influences of gender, race, degree-plan, and instruction. *Evolution: Education & Outreach, 11*(18), 1-14. https://doi.org/10.1186/s12052-019-0097-y

Schumacker, R. E. (2016). Reliability: A Rasch perspective. *Educational and Psychological Measurement, 67*(3), 394-409. https://doi.org/10.1177/0013164406294776

Sfendourakis, S. (2021). *In Darwin's mirror.* University Publications of Crete.

Smith, M. U. (2010). Current status of research in teaching and learning evolution: I. Philosophical/epistemological issues. *Science Education, 19*(6-8), 523-538. https://doi.org/10.1007/s11191-009-9215-5

Smith, Jr., E. V., Conrad, K. M., Chang, K., & Piazza, J. (2002). An introduction to rasch measurement for scale development and person assessment. *Journal of Nursing Measurement, 10*(3), 189-206. https://doi.org/10.1891/jnum.10.3.189.52562

Smith, M. U., Snyder, S. W., & Devereaux, R. S. (2016). The GAENE-generalized acceptance of evolution evaluation: Development of a new measure of evolution acceptance. *Journal of Research in Science Teaching, 53*(9), 1289-1315. https://doi.org/10.1002/tea.21328

Special Eurobarometer 225. (2014). Special Eurobarometer 225: Social values, science & technology (v1.00). *European Commission, Directorate-General for Communication*. http://data.europa.eu/88u/dataset/S448_63_1_EBS225

Stasinakis, P., & Athanasiou K. (2016). Investigating Greek biology teachers' attitudes towards evolution teaching with respect to their pedagogical content knowledge: Suggestions for their professional development. *EURASIA Journal of Mathematics, Science & Technology Education, 12*(6), 1605-161753. https://doi.org/10.12973/Eurasia.2016.1249a

Stasinakis, P., & Kampourakis, K. (2018). Evolution education in Greece. In D. Hasan, & L. A. Borgerding (Eds.), *Evolution education around the globe* (pp. 195-212). Spinger. https://doi.org/10.1007/978-3-319-90939-4_11

Stevens, S. S. (1946). On the theory of scales measurement. *Science, 103*(2684), 677-680. https://doi.org/10.1126/science.103.2684.677

Taber, K.S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education, 48*, 1273-1296. https://doi.org/10.1007/s11165-016-9602-2

Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis & Rheumatism (Arthritis Care & Research) 57*(8), 1358-1362. https://doi.org/10.1002/art.23108

Tormakangas, K. (2011). Advantages of the Rasch measurement model in analysing educational tests: An applicator's reflection. *Educational Research and Evaluation, 17*(5), 307-320. https://doi.org/10.1080/13803611.2011.630562

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Wu, H., & Leung, S.-O. (2017). Can Likert scales be treated as interval scales?–A simulation study. *Journal of Social Service Research, 0*(0), 1-6. https://doi.org/10.1080/01488376.2017.1329775