Evaluation of a Flipped Classroom Teachers Training Course Assessment Through Latent Trait Theory Analysis

Ioannis Katsenos, Spyros Papadakis, and George S. Androulakis

***Abstract*—Assessment of an educational program/course, based on quantitative data, is attempted in this study, by using the final deliverables of the trainees and assess them according to a predefined set of items connected to the desired Learning Outcomes and a predefined scale for each item. The statistical analysis of the items' grades, first using factor analysis and then using an Item Response Theory model, gives an indication of the Learning Outcomes' degree of achievement and consequently guides the training designers to modify training strategies for a potential next cycle of the training program/course. For this study, a teacher training course on flipped classroom methodology, has been used and the above concept was tested. Our analysis using Item Response Theory, revealed the Learning Outcomes partially or not at all achieved showing very good agreement with trainers' intuitive observations. For the future, the use of such a quantitative assessment could involve Structural Equation Modelling (SEM) tools to assess the relations among learning outcomes, prior knowledge and teaching practices and temporal analysis during training course execution using not only final data but also data from intermediate phases.**

***Index Terms*—Educational Assessment, Item Response Theory, Flipped Classroom.**

I. Introduction

It is common knowledge, that assessment of a training program is always at the last step of the program, to help the organizers understand what went good and what bad and whether the aims of the training program have been achieved. could be performed in various ways. It is also a common practice to have the trainees assess the program, they have just completed through a questionnaire. On the other hand, the trainers assess the achievement of the set educational outcomes for the trainees e.g. through exams or final projects etc. It is then on the trainers/teachers hand to intuitively combine the different aspects of training program assessment and using their experience to try to figure out what went good and needs to be retained and what wrong and needs to be payed attention during next cycle.

In this study, we attempt to provide a series of methodology steps, to help trainers/teachers analyze the final outcomes of a training program of any kind, by looking into the results of the learning outcomes evaluation of their trainees, using statistical tools/methodologies, i.e. here Exploratory Factor Analysis and Item Response Theory. The training program under analysis in this study, was a teacher training program on the flipped classroom teaching methodology, held by blended learning methodology over different western Greece's cities.

Published on September 23, 2019.

I. Katsenos is with the University of Patras/Department of Business Administration, Patras, Greece (e-mail: ikatsenos@gmail.com).

S. Papadakis is with Hellenic Open University, Aristotelous 18, Patra 263 35, Greece (e-mail: papspyr@gmail.com).

G. S. Androulakis is with the University of Patras/Department of Business Administration, Patras, Greece (e-mail: gandroul@upatras.gr)

II. Theoretical Framework

A. Latent trait models – Dichotomous Item Response theory models

In their simplest form unidimensional dichotomous latent trait models utilize a set of initially binary responses x_{i1}, x_{i2}, … x_{ik} to a set of k items (e.g. questions in an educational assessment test), for i examinees to calculate the probability of x_{ij} to be 0 or 1 given the true ability level θ_i of the examinees ($p(\theta_i) = P\{x_{ij} = 1|\theta_i\}$). To estimate $p(\theta_i)$, a monotonically increasing function in (-∞, +∞) for θ_i, most commonly the logistic function i.e. $p_{ij} = \Psi(\theta_i - \delta_j)$ is used. Depending on the number of parameters to include in the model 1,2,3 and 4-parameter models are derived. For instance, the 2-parameter logistic model [1]

$$P(x_{ij} = 1|\theta_i, a_j, \delta_j) = \frac{1}{1+e^{a_j(\delta_j - \theta_i)}} \quad (1)$$

for dichotomously answered items, calculates the probability of correct answering item j by someone with ability level θ_i given the item's difficulty δ_j (otherwise location parameter), and the item's discrimination parameter a_j.

The location parameter δ_j is that value of the ability θ at which there is 50% probability for the examining item to be correctly endorsed. Items with lower δ_j values are 'easier' and expected to be endorsed at lower trait levels. Item discrimination, denoted a_j for item j, describes how well an item can differentiate the examinees at different trait levels. In the simplest binary case, it is defined as the slope of the logit function at δ_j. The steeper the curve, the better the item can discriminate between entities with different levels of the trait.

Graphical representation of (1) results in the Item's Characteristic Curve (ICC)

IRT uni-dimensional models assume that 1) the probability of an entity selecting an item increases as the latent trait level increases (monotonicity) 2) all items equally contribute to the underlying latent trait (unidimensionality) 3) entity's trait level does not depend on which items are administered nor on the particular entities sample (invariance), 4) Responses are independent given an entity's ability level (local independence) and 5) The same item response functions applies to all members of the entities population.

Having determined the IRT parameters for all items, the ability level (IRT score) for each examinee, can be calculated, through different methods. For instance, according to maximum likelihood estimation procedure for 2-Parameter Logistic (2PL) model, starting from an initial arbitrary value, the probability of correct answers is calculated for each examining item according to equation (2) [2].

$$\widehat{\theta_{s+1}}^{\iota} = \widehat{\theta_s}^{\iota} + \frac{\sum_{i=1}^{N} a_i[u_i - P_i(\widehat{\theta_s})]}{\sum_{i=1}^{N} a_i{}^2[P_i(\widehat{\theta_s})Q_i(\widehat{\theta_s})]} \quad (2)$$

Where: θ_s is the ability estimation for s^{th} iteration, a_i is the discrimination parameter for i^{th} item, u_i is the response of the examined entity to item i (either 1 or 0), $P_i(\widehat{\theta_s})$is the probability of correct response to item i under the given model at ability θ and within iteration s, while $Q_i(\widehat{\theta_s})$is the corresponding wrong response probability.

The corresponding standard error for the measurement can be calculated by equation (3):

$$SE(\hat{\theta}) = \frac{1}{\sqrt{\sum_{i=1}^{N} \alpha_i{}^2 P(\hat{\theta}) Q(\hat{\theta})}} \quad (3)$$

The reciprocal of the squared $SE(\hat{\theta})$ for an item i is a function $I(\hat{\theta})$ defined as the Item Information Function (IIF), meaning that where $SE(\hat{\theta})$ is small, the value of θ is accurate and thus we obtain higher information than in areas where $SE(\hat{\theta})$ is higher. A test is a set of items, thus the Test Information Function (TIF) is the sum of all items' information at each ability level, or simply [2]:

$$I(\theta) = \sum_{i=1}^{N} I_i(\theta) \quad (4)$$

Where for instance for the 2PL model, $I_i(\theta) = a_i{}^2 P_i(\theta) Q_i(\theta)$ and a_i is the discrimination parameter for item i, $P_i(\theta) = \frac{1}{1+e^{-a_i(\theta-\delta_i)}}$, $Q_i = 1 - P_i$ and θ is the ability level

For each item, the information function has a peak near (or exactly at for 1PL and 2PL models) the difficulty δ_i and is symmetrical around it, while for the total test the total information function (TIF) may be flat over a range of θs. Graphical representation of $I_i(\theta)$ results in the Item's Information Curve (IIC), while for the whole test the respective graphical representation results in the Test Information Curve (TCC). IICs and TCC show graphically the ability ranges where Information is higher for each item and for the test. The ICC may be used for identifying the θ position of the maximum accuracy given by the item

B. Polytomous Item Response theory models – The Generalized Partial Credit Model (GPCM)

When the responses to items are polytomous (i.e. more than two, success-failure), e.g. 3 or more, the IRT models need to be extended. Masters [3] proposed the Partial Credit Model (PCM), in which modeling of the ordered polytomous data involves their decomposition into a series of ordered pairs of adjacent categories or category scores and then a dichotomous model is successively "applied" to each pair. The partial credit model specifies that the conditional probability that an examinee with latent location θ obtains a category score of x_j is

$$p(x_j|\theta, \delta_{jh}) = \frac{e^{\sum_{h=0}^{x_j}(\theta-\delta_{ih})}}{\sum_{k=0}^{m_j} e^{\sum_{h=0}^{k}(\theta-\delta_{ih})}} \quad (5)$$

Where δ_{jh} is the transition location parameter, which in effect, reflects the relative difficulty in endorsing category h over category (h – 1). The use of a subscript on m (i.e., m_j) reflects that the number of category scores may vary across items. Therefore, the Partial Credit Model may be applied to items that are polytomously scored with a varying number of category scores, are dichotomously scored, or consist of both dichotomously and polytomously scored items[4].

The probability of obtaining a particular category score as a function of θ may be graphically represented in an option response function (ORF); ORFs are sometimes referred to as category probability curves, category response functions, operating characteristic curves, or option characteristic curves.

Muraki in 1992 generalized the PCM in [5] by relaxing the assumption of equal discrimination parameters among items made by Masters (in other words, by extending the 2PL dichotomous model), thus proposing, that the probability of endorsing the k^{th} category x_{jk} of item j is given by

$$p(x_j|\theta, a_j, \delta_{jk}) = \frac{e^{\sum_{h=1}^{k_j} a_j(\theta-\delta_{ih})}}{\sum_{c=0}^{m_j} e^{\sum_{h=1}^{c} a_j(\theta-\delta_{ih})}} \quad (6)$$

Where θ is the latent trait, a_j is the item's discrimination, δ_{jh} is the transition location parameter between the h^{th} and the h-1 ctegory (i.e. the intersection point of adjacent ORFs), m_j is the number of categories for item j and $k=\{1,\ldots,m_j\}$. Muraki, arbitrarily defined the first boundary location as zero ($\delta_{j1}=0$), thus there are m_j-1 transition location parameters for each item with m_j categories m.

C. Learning outcomes' formulation and taxonomies

Assessment serves to diagnose, predict, place, evaluate, select, grade and guide students or teachers. That is, in all education levels, assessment results are used to decide about students (i.e., student advancement), to decide about teaching and learning (i.e., curriculum decisions) and increasingly assessments are linked with certification of competence and the validation of performance on job-related tasks[6]. While assessment can be seen as the Check step of a Plan-Do-Check-Act cycle in continuous learning process[7]–[10] formulation of the learning outcome objectives is the Plan step. Learning outcome objectives formulation, as well as the teaching methods to be used should be aligned to the learning activities assumed in the intended outcomes[11].

Several taxonomies have been proposed to formulate learning outcomes objectives, focusing on different aspects of learning processes. For instance, SOLO taxonomy proposed by Biggs[12], describes the increase in the ability of the trainee to associate principles to new ideas, while more recent Fink's taxonomy[13] is not hierarchical, but describes the intersection of six important to learning sections.

However, still the most popular taxonomy for formulating

learning outcomes remains revised Bloom's taxonomy[14]–[16] especially if only objectives at the cognitive level are to be pursuit.

D. Flipped classroom

Although there is no single definition, flipped classroom is generally characterized by its course structure comprising in-class and out-of-class activities. It uses classroom time for students to actively engage in interactive learning activities, while traditional lectures are delivered out of formal class time with videos, audios, content-rich websites, games and simulations[17]. Such a learning design intends to have classroom time to engage students in active learning [18] and to have the teacher a "guide on the side" instead of a "sage on the stage"[19]. Students are encouraged to explore and solve problems either independently or in groups collaborated to achieve their learning outcomes.

A review of the literature on the flipped classroom has discovered that there is still a pressing need for studying "how" to support teachers to design and implement the flipped classroom and, moreover, to be able to connect this pedagogical design with evidence of advantages related with various aspects of student learning.

III. Research questions and Methodology

A. Educational environment and assumptions for this study

Teacher's training in Greece is performed either through centrally designed and implemented programs by ministry of education, or by decentralized regional training centers - PEK- which are under restructuring after summer 2018. The Regional Training Center of Patras, at its last year of operation, organized a "Flipped Classroom methodology" training course spread over the Region of Western Greece.

The course involved 40 trainees and 376 trainers (class teachers - volunteers), in four groups (in seven different cities of western Greece), lasted 36 training hours (16h on site – 20h by distance), between February and June 2018. The methodology chosen and followed was blended learning, with sixteen hours physical presence in the classroom and twenty training hours asynchronous work assisted by LAMS platform[20] . The basic concept applied, was to deliver a course on Flipped Classroom methodology, using the same methodology as the course's training method. The technology supported used was the Learning Activity Management System (LAMS). The LAMS (https://www.lamsfoundation.org/) is the most widespread and popular platform that implements the ideas of learning The LAMS is an Online Free Open Source Software that supports the design, authoring, management and supervision of the execution of courses in the form of sequences of learning activities design [21], [22].

The core competence examining instrument for the trainees, was the composition of a learning scenario (LS), using the Flipped Classroom methodology. During the second week of their seminar, all the 376 participants were asked to design a learning scenario on a subject of their choice and implement it in their class using the FC model. The trainees began working on it in the fourth week and delivered it at the end of the course.

Each learning scenario was evaluated by a) the respective trainer and b) by other trainees of the same group according to a pre-specified measurement scale (see Appendix A). Three ordered category levels were defined for each formulated learning outcome.

The lower category level meaning was that the specified expected learning outcome was poorly or not at all achieved. The middle category level signified moderate learning outcome achievement, while the higher category level meaning was that the expected learning outcome was fully achieved.

B. Proposed methodology steps

The methodology steps followed in this study could be regarded as the steps to be followed for assessing any training course or even a training program assessment at any level.

1) Learning outcomes' formulation

As is common practice, the learning outcomes are formulated during course design and the learning activities are designed. Learning outcomes formulation can be done according to any taxonomy.

2) Learning outcomes' grading

It is done against a predefined set of items, assumed to assess the learning outcomes and using a preselected scale. It is done at the end of the training course/program by the trainers/teachers

3) Identification of underlying relations on the answers

Although a model for the evaluation items, might exist at their time of the formulation, these might have been perceived differently by the answering groups. Therefore, factor analysis (exploratory and/or confirmatory) is needed to unveil the underlying dimensions and facilitate polytomous unidimensional IRT analysis at the next step.

4) IRT analysis

The data sets containing the evaluation data are analyzed using an appropriately selected IRT model.

In IRT analysis studies, the examined entities (trainees) are in the rows and the examining items are in the columns of a table and this table is fitted according to an IRT model using an appropriate environment. Although there have been proposed ways to identifying the items' location for polytomous items[23] we expect that for the level of detail sought in applications like ours, simply the mean of location indices would be enough.

Another means for identifying the difficulty of the item is to check the θ value where information function $I(\theta)$ maximizes, this can be calculated or simply estimated by the relevant IIF graphs. Thus, each item could be characterized e.g. as low, medium or high difficulty if one divides θ scale to three regions

5) Learning outcomes evaluation

It is done by combining the information gathered from the previous steps

- The median of the grades for each item, which is associated to each learning outcome serves as an indication for deciding upon the achievement of this learning outcome. If the median of the grades of an item is e.g. into the lower answering category, one can safely conclude that the associated learning

outcome to this item has not been achieved. On the other hand, a median falling in the upper answering category, will imply the opposite.

- Dividing the ability level scale θ for each dimension to e.g. 3 categories (low, medium, high) and identifying the location of each item, one can conclude on the relative difficulty of each of them. Thus the items (and consequently the associated learning outcomes) could be characterized as e.g. lower, medium or high difficulty.

IV. Application and Results

A. *Learning outcomes' formulation*

It was done during the course design phase. The table presented in Appendix A, was produced and provided to the evaluators when learning scenario evaluations were requested. Revised Bloom's taxonomy was used in this study, however the taxonomy used is not expected to influence further analysis.

B. *Learning objectives grading*

In this study, each learning scenario was double evaluated by the teachers and peers, thus two sets of data were assembled for validity checking of the results, as described in previous section. Double evaluation should not however be necessary for normal application of this methodology and solely final evaluation of the trainees' outcomes by their trainers, should be enough.

C. *Identification of underlying relations on the answers*

Internal consistency of the data was examined, for Trainers' Dataset and for trainees' Dataset. Both datasets seem adequately internally consistent, giving Crombach's alpha factor values 0.73 and 0.79 respectively.

Exploratory factor analysis revealed four factors for the trainers' dataset and three underlying factors for the trainees' dataset, as shown in Table I.

In both cases there exist two evaluation items (LGF1 and CAD5 for trainers' dataset and TOO2 & VD1 for trainees' dataset), which seem not to be related enough to the overall data structure therefore they were excluded from further analysis.

Confirmatory factor analysis for the above identified models, gave good results and confirmed model selection, as shown in Table II. Factor analysis reveals 3 dimensions for trainees' dataset and four dimensions for trainers' data set, which correspond to different latent abilities identified. One can observe that:

- Dimension 1 corresponds to factor F1 with item CE5 mainly contributing by both datatsets.
- Dimension 2 is mainly formed by items VC4, VA5 and CL3 in both datasets (factors F3 in both datasets)
- Dimension 3 is mainly formed by items VO6, VI2 and PK2 (Factor F2 in trainees' dataset and Factor F4 in trainers' dataset)
- Dimension 4 exists only into trainers' dataset consisting by items TOO2, CAT3 and COO3

D. *IRT analysis*

In this study, the results of the evaluations were gathered electronically and analyzed using the polytomous IRT GPCM model.

The R environment[24] and the mirt package[25] were preferred for IRT analyzing the data, however similar results are expected whatever IRT analysis tools are used.

The IRT analysis was performed per identified dimension and the parameters calculated were:

1. The discrimination parameter (a_i)
2. The location parameters (thresholds, δ_{1i} and δ_{2i})
3. The total information area under item information curves (as a single measure of the total accuracy in the items measurement)

TABLE I: FACTORS' STRUCTURES

Underlying model for trainers' dataset	F1=~ SE5 + CE5 + VD1 + VS4 F2=~ TOO2 + CAT3 + COO3 F3=~ CL3 + VA5 + VC4 F4=~ VO6 + VI2 + PK2
Underlying model for trainees' dataset	F1=~ CE5 + SE5 + CAD5 + CAT3 F2=~ VS4 + VO6 + VI2 + PK2 F3=~ VC4 + VA5 + LGF1 + CL3 + COO3

TABLE II: ITEMS' LOADINGS ON FACTORS AND DIMENSIONS (ABILITIES) IDENTIFIED

Trainers' Evaluation				
N=74				
X2=55.332, df=59, p=0.611				
CFI=1.000, TLI=1.041				
RMSEA=0.000, SRMR=0.082				
	Estimate	Std.Err	z-value	P(>|z|)
F1 =~				
SE5	1.000			
CE5	1.212	0.324	3.743	0.000
VD1	-0.211	0.092	-2.295	0.022
VS4	0.658	0.199	3.299	0.001
	Estimate	Std.Err	z-value	P(>|z|)
F2 =~				
TOO2	1.000			
CAT3	1.950	0.747	2.612	0.009
COO3	0.758	0.241	3.152	0.002
	Estimate	Std.Err	z-value	P(>|z|)
F3 =~				
CL3	1.000			
VA5	1.254	0.429	2.919	0.004
VC4	0.986	0.345	2.858	0.004
	Estimate	Std.Err	z-value	P(>|z|)
F4 =~				
VO6	1.000			
VI2	1.143	0.405	2.825	0.005
PK2	0.672	0.239	2.810	0.005
Trainees' Evaluation				
N=135				
X2=67.619, df=62, p=0.291				
CFI=0.984, TLI=0.980				
RMSEA = 0.026, SRMR=0.059				
	Estimate	Std.Err	z-value	P(>|z|)
F1 =~				
CE5	1.000			
SE5	0.748	0.104	7.193	0.000
CAD5	0.655	0.096	6.850	0.000
CAT3	0.458	0.106	4.338	0.000
	Estimate	Std.Err	z-value	P(>|z|)
F2 =~				
VS4	1.000			
VO6	0.582	0.104	5.612	0.000
VI2	0.866	0.145	5.990	0.000
PK2	0.767	0.135	5.700	0.000
	Estimate	Std.Err	z-value	P(>|z|)
F3 =~				

VC4	1.000			
VA5	0.643	0.149	4.300	0.000
LGF1	0.687	0.179	3.847	0.000
CL3	0.358	0.136	2.641	0.008
COO3	1.425	0.255	5.587	0.000

1) Dimension 1

As can be seen in Table III and Fig. 1,

- The item VD1 provides no information as almost all answers fall in the third answering category (i.e. the video duration is short).

TABLE III: IRT PARAMETERS FOR DIMENSION 1

	Trainers' Dataset			Trainees' Dataset		
Item	a_i	δ_{1i}	δ_{2i}	a_i	δ_{1i}	δ_{2i}
SE5	2.91	-1.33	0.14	2.36	-1.43	-0.40
CE5	1.53	0.35	0.53	2.67	-0.59	0.04
VD1	**0.03**	**-39.9**	**-117.74**			
VS4	0.94	1.39	2.09			
CAD5				1.76	-1.81	-0.45
CAT3				0.59	-2.11	-0.56

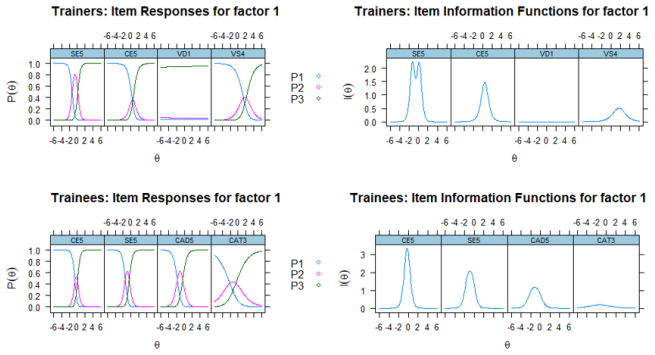

Fig. 1. Item Response Curves and Item Information Curves for Dimension 1

TABLE IV: IRT PARAMETERS FOR DIMENSION 2

	Trainers' Dataset			Trainees' Dataset		
Item	a_i	δ_{1i}	δ_{2i}	a_i	δ_{1i}	δ_{2i}
CL3	2.09	-2.80	-1.0	0.68	-2.94	-2.5
VA5	1.96	-2.24	-0.52	1.36	-2.72	-0.90
VC4	1.16	-3.15	-0.64	2.32	-2.19	-0.40
LGF1	-	-	-	0.73	-2.62	-0.62
COO3	-	-	-	1.24	-0.77	-0.20

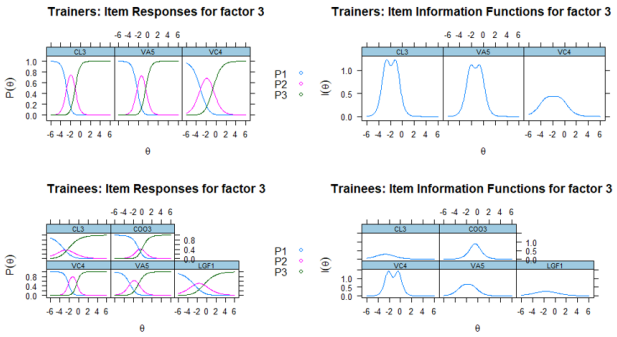

Fig. 2. Item Response Curves and Item Information Curves for Dimension 2

- The items SE5 and CE5 have similar characteristics, with thresholds in the middle of θ scale and similar amount of information provided.
- The item VS4, being important only in trainers' dataset, discriminates better examinees with higher ability level, while items CAD5 & CAT3, being
- Important only to trainees' dataset, discriminate better examines with lower θ levels.

2) Dimension 2

As can be seen in Table IV and Fig. 2, in dimension 2 the items providing information are different among the two datasets, so while in the trainers' dataset item CL3 has the highest discrimination and therefore provides the maximum information, in the trainees' dataset the highest discrimination comes from item VC4. Though, dimension 2 items for both datasets measure ability for θs below zero.

3) Dimension 3

As can be seen in Table V and Fig. 3,

- Item PK2 for the trainers' dataset doesn't contain any answer from the highest-level category and therefore no high threshold δ_{2i} or information function can be computed. Also, for the trainee's dataset, this item indicates that this learning objective has been proven difficult to achieve for the examinees as $\delta_{2i}>\delta_{1i}$ indicating that the answering probability of the second category level is never higher than the first and the second, thus the evaluators tend to choose more likely between the first and the third answer category.
- Item VO6 is the dominant for both datasets, providing most of the information and examining higher θs. Item VS4, for the trainees' dataset also functions well at high θs. Consequently, this dimension seems to examine higher θs

4) Dimension 4

Dimension is only present to the trainers' dataset and provides most accurate results in the mid-lower range of θ ability. As seen in Table VI and Fig. 4, item CAT3, has very high discrimination and provides high information between δ_1=-1,123 and δ_2=0,386 i.e. mid-low range of θs.

TABLE V: IRT PARAMETERS FOR DIMENSION 3

	Trainers' Dataset			Trainees' Dataset		
Item	a_i	δ_{1i}	δ_{2i}	a_i	δ_{1i}	δ_{2i}
VO6	2.00	1.02	2.08	1.73	0.60	2.38
VI2	1.68	-0.93	1.12	-	-	-
PK2	0.61	1.15	-	0.95	1.36	1.12
VS4	-	-	-	2.34	0.29	1.04
TOO2	-	-	-	0.52	-1.60	1.73

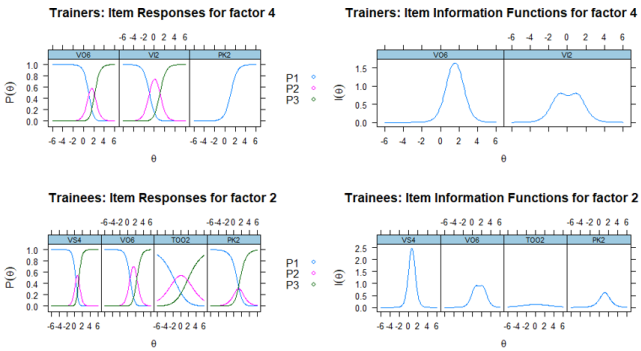

Fig. 3. Item Response Curves and Item Information Curves for Dimension 3

TABLE VI: IRT PARAMETERS FOR DIMENSION 4

Factor F2 for Trainers - no such dimension for trainees dataset						
	Trainers' Dataset			Trainees' Dataset		
Item	a_i	δ_{1i}	δ_{2i}	a_i	δ_{1i}	δ_{2i}
TOO2	0.95	-0.54	0.81	-	-	-
CAT3	0.255	-1.12	0.39	-	-	-
COO3	1.08	-2.38	0.36	-	-	-

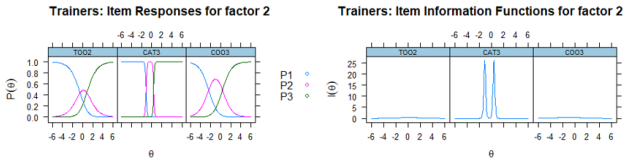

Fig. 4: Item Response Curves and Item Information Curves for Dimension 4

E. Learning outcomes evaluation

The majority of evaluations, as indicated by the median of each dataset, can be used to conclude on the level of achievement for each learning objective.

In this study, we had double learning outcomes evaluation, thus a temporal evolution of the learning scenario development may be observed (could be further investigated in a future study) in some cases, since the examinees had the opportunity to work further onto their deliverables and further evolve them.

TABLE VII: LEARNING OUTCOMES EVALUATION

Item	Median Trainers	Median Trainers	Ability level	Learning outcome estimate	Remarks
LGF1	1	2	Lower	Achieved	Improved
VD1	2	2	Lower	Achieved	
TOO2	1	1	Mid	Partially achieved	
PK2	0	0	Higher	Not Achieved	
VI2	1	1	Mid	Partially Achieved	
CL3	2	2	Lower	Achieved	
CAT3	1	2	Lower	Achieved	Improved
COO3	1	1	Lower	Partially Achieved	
VC4	2	2	Higher	Achieved	
VS4	0	0	Higher	Not Achieved	
VA5	2	2	Lower	Achieved	
CAD5	2	2	Lower	Achieved	
SE5	1	2	Mid	Achieved	Improved
CE5	0	1	Mid	Achieved	Improved
VO6	0	0	Higher	Not Achieved	

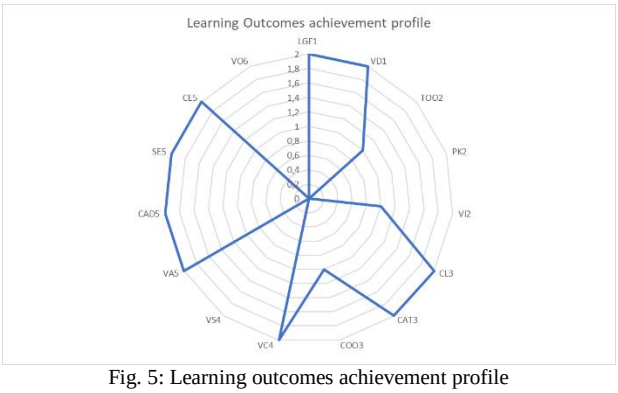

Fig. 5: Learning outcomes achievement profile

Moreover, based on the point in ability level θ, where occurs the maximum of the information function I(θ) (as roughly identified by the relevant figures above), each item can be characterized as of lower, medium or higher ability level.

Thus, we can summarize the IRT analysis results on Table VII, which provides a picture of the Learning Outcomes examined by each item and its degree of achievement.

Furthermore, Fig. 5, may be constructed to visualize the level of achievement for the different learning outcomes of each particular cycle of the training course or different training courses, based on the information presented in Table VII.

V. DISCUSSION AND FURTHER INVESTIGATION

Factor analysis showed that initial learning outcomes classification in accordance to revised Bloom's taxonomy was not accurate. Factor Analysis identified less than 6 dimensions in the data, which implies that the evaluators perceived a different interrelation among the learning outcomes and some of the levels are merged.

Teachers were able to identify one more dimension although they evaluated at earlier stage and the examinees continued to develop the learning scenarios.

For some of the evaluating items, there was further evolution (e.g. LGF1, SE3) between first evaluation by trainers and the second one by the trainees themselves.

Nine out of fifteen learning objectives seem to have been finally achieved (LGF1, VD1, CL3, CAT3, VC4, VA5, CAD5, SE5, CE4), however only two of them (CE5, SE5) seem to correspond to middle level ability θ.

Three out of fifteen learning objectives seem to have been partially achieved, with two of them having been examining middle layer abilities. There also exist 3 out of fifteen learning objectives which have not been achieved, all of them seeming difficult for the examinees.

The above results, indicate that the trainees only partially applied the instructed theory in for formulating their goals, for adding interactivity to their videos and for including group activities in their flipped classroom educational scenarios and that also they completely failed to use original videos or videos with high modification grade, they didn't have adequate interactivity to their videos and they also failed to specify correctly the prior knowledge into their scenarios.

The above observations coming only from the quantitative analysis of the learning scenario deliverables, agreed to a high degree to the observation of the trainers and constitute the basic issues to be addressed to a potential next cycle of this teacher training course by designing the appropriate training strategies.

Regarding the research questions of this study, it is evident that quantitative assessment of the deliverables of a training program/course can lead to conclusions regarding its effectiveness and guide onto which areas should be treated with special care in a next possible cycle of the educational program/course.

Further investigation could be done to allow increase of the depth of the quantitative analysis, by assessing learning outcomes achievement at both at intermediate and final phases. Also, the use of Structural Equation Modelling

(SEM) tools to assess the relations among learning outcomes, prior knowledge and teaching practices could lead to further deepening of the quantitative analysis, thus enable a deeper view of any training program.

APPENDIX A

Bloom's revised taxonomy levels	**Learning outcome**	**Evaluation categories**		
Remember	Learning Goals Formulation (LGF1)	1	2	3
		All used verbs are generic (to know, to learn)	There are some generic but also some specific verbs (to count, to describe)	All verbs are specific (to compare, to formulate)
	Video duration (VD1)	1	2	3
		The video is very long (>16´)	The video is long (10-15´)	The video is short (1-9´)
Understand	Type of objectives in the Learning Scenario (TOO2)	1	2	3
		Gain new knowledge (remember, understand)	Deepening knowledge (apply – analyze)	Create knowledge (evaluate – create)
	Prerequisite knowledge (PK2)	1	2	3
		The learning scenario does not mention any prerequisite knowledge	The learning scenario just mentions the prerequisite knowledge	The learning scenario just mentions the prerequisite knowledge and they are assessed with embedded questions in the video
	Video interactivity (VI2)	1	2	3
		The video does not contain or involve any interactivity of any kind (links or questions to be answered by the student)	The video contains or is followed by a few and elementary interactivity elements	The video has high interactivity
Apply	Learning Outcomes number sought – relation to course length (CL3)	1	2	3
		Too many learning outcomes are sought, not feasible to achieve all of them in the designed course length	Number of learning outcomes might be high, there is a risk that not all of them will be achieved	Number of learning outcomes sought is feasible in the designed course length
	Classroom activities types (CAT3)	1	2	3
		The designed classroom activities focus on remembering and understanding	The designed classroom activities focus on analyze and apply	The designed classroom activities focus on evaluation and creation fo new knowledge
	Cooperation (COO3)	`1	2	3
		All the designed classroom activities are to be executed individually	There are some activities to be performed individually, but there are also group activities designed	All designed activities are to be performed in groups and promote cooperation
Analyze	Video Content (VC4)	1	2	3
		The video content is poor (poor graphics and sound quality)	The video content is moderately appealing (acceptable graphics quality and good sound)	The video content is highly appealing (high quality graphics and crystal clear sound)
	Video supervision (VS4)	1	2	3
		The teaches has no supervision on the video sent to his/her students	The learning scenario mentions some supervision to the video by the teacher	The learning scenario contains specific field for the teacher to note his/her remarks from the video supervision, which is completed before the course
Evaluate	Video adequacy (VA4)	1	2	3
		The selected video serves too few of the intended learning outcomes	The selected video serves many but not all the intended learning outcomes	The selected video serves many but **all** the intended learning outcomes
	Classroom Activities description (CAD5)	1	2	3
		The description of the classroom activities contain only their titles	The description of the classroom activities is short	There is a complete description of the classroom activities and there are also Activity Sheets provided, where needed.
	Student Evaluation (SE5)	1	2	3
		The learning scenario does not contain students' evaluation activities	There are some evaluating activities designed	The learning scenario contains full description of evaluating activities for all intended learning outcomes
	Course Assessment (CE5)	1	2	3
		The learning scenario does not foresee any course evaluation	The learning scenario foresees partial course evaluation	The learning scenario foresees full evaluation of the learning outcomes achievement, the video and the classroom activities
Create	Video Origin (VO6)	1	2	3
		The video was found on the WWW and used by the trainee without any processing	The video was found on the WWW and was adapted by the trainee (e.g. narration, added comments, added questions)	The video was created from scratch and processed by the trainee

REFERENCES

[1] W. Revelle, "The ' New Psychometrics ' – Item Response Theory," *An Introd. to Psychom. theory with Appl. R*, pp. 241–269, 2010.
[2] F. B. Baker, *Item Response Theory the Basics of Item Response Theory*. ERIC, 2001.
[3] G. N. Masters, "A RASCH MODEL FOR PARTIAL CREDIT SCORING," 1982.
[4] J. de A. R., *The Theory and Practice of Item Response Theory*. New York: The Guilford Press, 2009.
[5] E. Muraki, "A GENERALIZED PARTIAL CREDIT MODEL: APPLICATION OF AN EM ALGORITHM," 1992.
[6] I. Lamprianou and J. A. Athanasou, *A Teacher's Guide to Educational Assessment*, vol. 13, no. 1. 2009.
[7] D. J. Flinders and S. J. Thornton, *The Curriculum Studies Reader Fourth Edition Edited by*. 2013.
[8] R. Moen and C. Norman, "Evolution of the PDCA Cycle."
[9] "Quality cycle - Eqavet." [Online]. Available: https://www.eqavet.eu/eu-quality-assurance/glossary/quality-cycle. [Accessed: 11-Nov-2018].
[10] M. Tam, "Outcomes-based approach to quality assessment and curriculum improvement in higher education," *Qual. Assur. Educ.*, vol. 22, no. 2, pp. 158–168, 2014.
[11] J. Biggs, "ALIGNING THE CURRICULUM TO PROMOTE GOOD LEARNING." 2002.
[12] J. Biggs, "INDIVIDUAL DIFFERENCES IN STUDY PROCESSES AND THE QUALITY OF LEARNING OUTCOMES," 1979.
[13] L. D. Fink, "A Self-Directed Guide to Designing Courses for Significant Learning." [Online]. Available: http://www.bu.edu/sph/files/2011/06/selfdirected1.pdf. [Accessed: 11-Nov-2018].
[14] L. W. Anderson *et al.*, *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives, abridged edition*, Abridged e. N.Y., 2001.
[15] M. Forehand, "Bloom's Taxonomy from Emerging Perspectives on Learning, Teaching and Technology," *Int. J. Educ. Manag.*, vol. 26, no. 2, pp. 205–222, 2012.
[16] D. R. Krathwohl, "A revision of bloom's taxonomy: An overview," *Theory Pract.*, vol. 41, no. 4, pp. 212–218, 2002.
[17] P. Baepler, J. Walker, M. D.-C. & Education, and undefined 2014, "It's not about seat time: Blending, flipping, and efficiency in active learning classrooms," *Elsevier*.
[18] B. Sohrabi and H. Iraj, "Implementing flipped classroom using digital media: A comparison of two demographically different groups perceptions," *Comput. Human Behav.*, vol. 60, pp. 514–524, Jul. 2016.
[19] A. King, "From Sage on the Stage to Guide on the Side," *Coll. Teach.*, vol. 41, no. 1, pp. 30–35, Jan. 1993.
[20] J. Dalziel, "Interact Integrate," in *20th Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education*, 2003, no. December, pp. 7–10.
[21] J. Dalziel, "Implementing learning lesign: The Learning Activity Management System (LAMS)," in *Proceedings of the 20th Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education (ASCILITE)*, 2003, pp. 593–596.
[22] S. Papadakis, N. Dovros, G. Paschalis, and E. Rossiou, "INTEGRATING LMSs IN THE EDUCATIONAL PROCESS:Greek Teachers' Initial Perceptions about LAMS," *Turkish online J. distance Educ. TOJDE*, vol. 13, no. 4, p. 55, 2000.
[23] U. S. Ali, H.-H. Chang, and C. J. Anderson, "Location Indices for Ordinal Polytomous Items Based on Item Response Theory," *ETS Res. Rep. Ser.*, vol. 2015, no. 2, pp. 1–13, Jun. 2015.
[24] RCoreTeam, *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2013.
[25] R. P. Chalmers, "mirt: A multidimensional item response theory package for the R environment," *J. Stat. Softw.*, vol. 48, no. 6, pp. 1–29, 2012.

Ioannis Katsenos has been born in Patras, Greece in July 1967. He got his Degree in Physics from the Department of Physics of the University of Patras, Greece in 1991 and one year later he was awarded an M.Sc. in Microwave Solid State Physics by the University of Portsmouth, UK
He completed his military service and after a period of three years he has been working as a freelancer teacher (teaching physics as well as informatics) in different organizations, he was recruited by Intracom S.A. as software engineer in the field of real time telecommunications software development. He stayed with Intracom S.A. for about seven years assuming also positions of software quality assurance manager, software project manager and line/project office manager. He switched to Computer Technology Institute & Press Diophantus (former Research Academic Computer Technology Institute), in Patras and worked for almost three years in educational software projects (development and procurement management) for the Greek Ministry of Education. He is now working as a physics teacher in a senior high school of Patras (Geniko Lykeio Kastritsiou), after nationwide selection process by Greek Ministry of Education in 2006. He continues to collaborate with Computer Technology Institute & Press Diophantus in Teacher Training projects. Since 2015, he is a Ph.D. candidate in Business Administration Department of University of Patras, under the supervision of Associate Professor George S. Androulakis.

Spyros Papadakis has been born in Chania – Crete, Greece in December 1961 He holds a PhD in Computer Science and Information Systems from the School of Science and Technology of the Hellenic Open University (HOU), a Master's Degree (Med) in Adult Education (HOU), a Postgraduate Certificate (PGCE) in Open and Distance Education (HOU) and a Bachelor's Degree (B.Sc.) in Mathematics from the University of Patras, Greece.
He is Organizational Coordinator, Regional Centre for Educational Planning in Western Greece Educational Coordinator, Computer Science, Western Greece Adjunct Faculty, Hellenic Open University (HOU). His research involves human-centered design of advanced technologies for learning, learning design, distance and blended education and training. It involves gaining a deep understanding of how people interact, collaborate, work, and learn as a foundation for the design of novel e-learning, mobile learning and blended learning systems. Spyros Papadakis is a member of the editorial board of one international journal and he serves as a reviewer for journals and conferences.

George S. Androulakis, born in 1969 was awarded his B.Sc. and M.Sc. degrees from the Department of Mathematics, University of Patras, Greece, where he also received his Ph.D. in the Unconstrained Optimization Methods. Having worked as a tutor at the departments of Mathematics, Pharmacy and Geology at the University of Patras currently is an Associate Professor in the field of Quantitative Methods at the department of Business Administration; he is also teaching at the postgraduate course in Mathematics for Decision Making, University of Patras. His interests are focused in nonlinear unconstrained optimization, neural networks training, optimal Runge-Kutta methods, systems of non-algebraic and/or transcendental functions.