Transnational Comparative Assessments in European Higher Education

PHYSICS

Measuring and Comparing Achievements of Learning Outcomes in Higher Education in Europe
2023

Measuring and Comparing Achievements of Learning Outcomes in Higher Education in Europe
CALOHEE Phase 2

The CALOHEE Phase 2 Project has been supported by the European Commission through the Erasmus+ Programme, Action Forward Looking Projects, 2020-2023.
Project number: 2019-612892

Tuning Educational Structures in the World

The name TUNING was chosen for higher education projects and initiatives to reflect the idea that universities do not look for uniformity in their degree programmes or any sort of unified, prescriptive or definitive curricula but simply for points of reference, convergence and common understanding. The protection of the rich diversity of higher education in Europe and the world has been paramount in the Tuning initiative from its start in 2001 and in no way seeks to restrict the independence of academic and subject specialists, or undermine local and national academic authority.

Co-funded by
the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

Publisher: International Tuning Academy, Universities of Deusto and Groningen
Bilbao and Groningen, 2023

Transnational Comparative Assessments in European Higher Education

PHYSICS

Marta Carli
Ornella Pantano, eds.

Measuring and Comparing Achievements of Learning Outcomes in Higher Education in Europe
2023

Table of Contents

Introduction

The context of higher education has been changing during the last 25 years, as a result of rapid advances in digitalization and methods of communication, job market disruption, politics and recently COVID-19, disruptive conflicts and inflation. The need for change of higher education learning has become even more imperative. Awareness of these challenges go back to the 1990s and resulted in EU initiatives and the Sorbonne/Bologna Declarations. This led to the call for developing a European Higher Education Area (EHEA).

A cornerstone of developing a EHEA is trust and confidence. The Area was launched in the context of the Bologna Process. This was thought necessary to enhance the quality and relevance of higher education for individual development, employment opportunities, societal needs. Another aspect was and is to have instruments in place to facilitate large scale credit mobility and recognition. Towards this end four key instruments have been developed: the *European Standards and Guidelines for Quality Assurance*, the *European Credit Transfer and Accumulation System* and the *Lisbon Recognition Convention* as well as two parallel and overlapping qualifications frameworks, the *Qualifications Framework for the European Higher Education Area* (QF for the EHEA) and the *European Qualifications Framework for Lifelong Learning* (EQF). The first defined in the context of the Bologna Process and the second initiated by the European Commission. Both have been endorsed by national authorities.

Qualifications frameworks are the foundations of the other instruments. They offer the reference point for the academic structure (curriculum design and credentials), quality assurance and accreditation as well as recognition of (period of) studies. Qualifications Frameworks encompass all three cycles of higher education learning.

In parallel, two major initiatives were taken, namely, the development of the QAA-UK Benchmark papers and the *Tuning Guidelines and Reference points* at subject area (discipline) level. These proved to be pivotal for giving substance to develop and enhance degrees and to move from expert driven education toward student-centred and active learning. Both initiatives were developed by groups of academics, however, many academics have found it difficult to deal with this fundamental change of the learning paradigm. Lack of initial training and continuing professional development have continued to hinder large scale change. This has been exacerbated by the over-complex structures in place. That is having two European overarching frameworks and subject ones which are not fully aligned. This might have drained away full adoption of the instruments available.

To respond to this concern, a proposal has been made by the Tuning initiative, called *Measuring and Comparing Achievements of Learning Outcomes in Europe* (CALOHEE), to make a deep analysis of the strength and weaknesses of the existing models. This has resulted in *General Tuning-CALOHEE Qualifications Reference Frameworks* for all three cycles, as well as aligned reference frameworks on the level of subject areas. An important driver for developing these frameworks has been to make the implicit explicit.

These much more detailed frameworks, building on the existing ones, offer the opportunity to encompass present and future challenges. In addition, ten subject areas have been, and are, developing Subject Area Learning Outcomes Reference Frameworks. These offer a template and menu as to what can be learned in the context of a degree programme.

This resulting set of reference frameworks will reduce complexity, offer greater clarity and guidance for programme design, delivery and quality assurance.

However, qualifications reference frameworks are only part of process of change. As fundamental and as a consequence of the change of the paradigm of learning, is revisiting the way learning, teaching and assessment is designed and undertaken. This has been done too in the context of the CALOHEE initiative, supported by the European Commission.

Preparing international comparative assessments

Mutual recognition and mobility go hand in hand and therefore need evidence of comparability of learning and teaching, but in particular assessment, which should obviously be aligned.

Although General Qualifications Reference Frameworks, Subject Area Qualifications Frameworks and related Subject Area Learning Outcomes / Assessment Reference Frameworks offer clarity regarding the levels of learning, they do not offer the evidence whether the related learning is actually achieved. To achieve the latter some form of assessment must take place, primarily to assure that across the spectrum of countries and institutions comparable learning in terms of its outcomes is taking place.

On the level of achievement, it is possible to make a distinction between the individual learner, the subject, the programme, the HE institution and the country (system level). The aim of the CALOHEE project has been to develop diagnostic international comparative assessments for five disciplinary fields, that is civil engineering, history, nursing, physics and teacher education.

These assessments provide a diagnostic tool to allow for a comparison to be made regarding the level of achievements of the different descriptors as included in the frameworks. The focus is here on the degree programmes in the context of the subject area. The results of the exercise will provide valuable evidence-based information for academic staff responsible for delivering the programme to allow for further enhancement.

The discussions among international groups of subject area experts show us that disciplines have their own requirements. There are obviously specific contextual settings, cultural and national conditions. For example, the field of history only allows for a high level of abstraction, whereas nursing, civil engineering and teacher education are usually regulated professions with all that that entails.

Assessment of students is perceived as a highly sensitive issue and the prime responsibility of the academic. However, academics are together responsible for implementing a programme. This requires coordination regarding programme design, delivery, evaluation and student-assessment and grading. This does not touch academic freedom. Although all programmes will have their own profile, there should be common standards meeting international reference points. This approach intends to do justice to the EU motto, introduced in 2000, 'unity in diversity' which is clearly not standardisation.

In this context, the relation should be highlighted between the graduate profile and the learning outcomes of an individual programme and its units. This reflects the different missions of institutions and programmes, covering the full spectrum from research driven programmes to applied ones. This can be visualised in a spider web in which individual degree profiles, programme and unit learning outcomes are matched with the CALOHEE subject area qualifications refence frameworks for all three cycles, representing the graduate profile. These spiderwebs show varieties, which are both system and programme related.

Regarding the system level, although pursuing the EHEA, it has to be fully understood that we are dealing with national states which historically have their own educational philosophies, cultures and traditions. Regarding philosophies we can make a distinction between the Anglo-Saxon, Humboldtian, Napoleonic and Soviet models. These traditions are deeply rooted and have an ongoing impact on the way learning, teaching and assessments is constituted, although convergence is taking place. This convergence – implying international alignment at subject area / disciplinary level - is commended by global societal developments and needs, to which the higher education sector and its degree programmes are expected to respond.

At programme level, countries might still define conditions which have to be met and/or set limits regarding the autonomy of the professional. This has implications for the (transnational) assessments to design.

As a consequence, in valid transnational comparative assessment both communalities and differences should be taken into account, as they have been detailed above. In this setting, lessons have been learned from the OECD *Assessment of Higher Education Learning Outcomes* (AHELO) feasibility study, implemented in the period 2010-2013, which obtained severe criticism from policy makers as well as academics, because it did insufficiently recognise the wide range of system and programme differentiations.

The disciplinary experts, involved in this CALOHEE project, are fully aware of the diversity in the way learning, teaching and assessment is modeled, although at the same time agreeing on the descriptors as defined in their subject area qualifications reference frameworks and far more detailed learning outcomes / assessment reference frameworks. Finding common ground - doing justice to the differences - has taken considerable time, but proved to be conditional for developing useful (transnational) assessments.

Departing from the objectives of the Bologna Process and the EHEA that programmes should be outcome based, the assessments developed, intend to cover high level generic and subject specific competences, that is applying knowledge and skills in real life situations – work place and society – requiring 'autonomy' and 'authority'. Authority reflecting self-confidence to take position and act accordingly. In other words, the assessments should allow for evidencing a critical mindset in the context of a particular academic field by focussing on 'measuring' high level skills and competences in the context of the subject area and its domain of knowledge, such as critical thinking, analyzing and synthesizing, making and criticizing an argument, problem solving, observing and analyzing behavior, operating in conjunction with others. All perceived from two angles: the academic field involved and active societal participation. Relating to present and future needs of society, a much wider scope and approach than 'disciplinary knowledge and skills' and 'critical thinking' as had been tested in the global OECD-AHELO feasibility study.

This requires taking into account 'burning societal issues', for which in the context of the CALOHEE projects separate initial reference qualifications frameworks were prepared, meant to serve as sources of information and inspiration. Based on academic literature and policy documents, it identified five current topical issues, that is:

- Societies and Cultures: Interculturalism
- Processes of information and communication
- Processes of governance and decision making
- Ethics, norms, values and professional standards
- Sustainable development (climate change)

These topical issues should be integrated in the actual learning, teaching and assessment processes doing justice to the academic field involved and avoiding overload of learning.

From the start of the CALOHEE project to develop transnational assessments and testing, the aim has been mutual. The outcomes should allow for real testing to be applicable in different contexts, ranging from an individual HE education programme to transnational testing. Intended to be inspirational – offering new models of assessment – they should also be aspirational by covering topical issues.

As has been indicated already a distinction is made between the development of models of assessment and actual assessments and testing. Testing is defined here as the application of the assessments prepared, by asking groups of students to take a test. According to the project aim, actual testing was not foreseen in this phase. This project focussed instead on preparing the groundwork for testing whether of theory or in the workplace where this is relevant for the programme.

In the context of the CALOHEE Phase 2 project assessment models and assessments have been prepared for the following five subject areas: Civil Engineering, History, Nursing, Physics and Teacher Education, nearly covering the full range of academic fields.

The assessments have been developed to measure the achievements of generic and subject specific competences at the end of the bachelor / first cycle.

Structure of the assessments

The five subject area groups have followed a comparable model and approach to implement their tasks. Due to the COVID-19 pandemic initially the meetings took place online. Because more fundamental discussions were needed to define common ground requiring deep intensive reflection over a longer time span, only limited results could be obtained. Three multi day face-to-face meetings were needed to come up with actual results. These meetings took place in the period April – September 2022 and were supported by an additional set of online meetings.

A first step has been to match individual degree programmes with the subject area qualifications reference framework published in 2018. A follow-up has been to re-visit their academic field making use of the 2018 edition of the brochure *Tuning Guidelines and Reference Points for the Design and Delivery of Degree Programmes* for their subject area. This proved to be a learning process in itself, developing partly new insights requiring accommodations of the materials prepared earlier.

The third step was to identify the (sub) descriptors included in the qualifications reference framework and learning outcomes / assessment reference framework, best suitable for developing transnational assessments, but also key to the subject area. This again required fundamental and deep reflections. The next step was to identify the most appropriate mode(s) of assessment and to decide on its feasibility. Independently of the mode of teaching and learning - class room, online, hybrid - different assessment formats were suggested to apply, e.g. scenario testing, observation, critically responding to arguments / texts, analyzing a problem and coming up with possible solutions, etc. This to be followed by describing / documenting the overview of items and approaches (independent of existing individual degree programmes) and the choices made. In practice, to:

- identify for each of these items the modalities for assessment: learning/teaching required, the best ways of assessment and the criteria for assessment.
- document the rational for selecting a particular competence; describe the actual test
- constitute a set of assessments reflecting a key part of the descriptors as included in the qualifications reference framework. The result should be a variety of assessment formats for the competences identified.

The outcomes of the work established by the five subject area groups are presented in separate publications for each of the five subject areas involved in the CALOHEE Phase 2 project: Civil Engineering, History, Nursing, Physics and Teacher Education. The reports of these five disciplinary groups follow a comparable format, but each group has taken the freedom to make its own choices in presenting its findings in doing justice to the process of reflection and discussion. In this brochure is presented the work established by the Subject Area Group of Physics, coordinated by dr. Ornella Pantano with substantial support of dr. Marta Carli, both affiliated with the University of Padova, Italy.

CALOHEE Project Team
Groningen, 2023

0. The Tuning–CALOHE2 Physics Subject Area Group (2020-2022)

Coordinator	
Italy Ornella Panatano Università degli Studi di Padova / University of Padova (UNIPD) ornella.pantano@unipd.it	
Members	
Belgium Katrien Strubbe Universiteit Gent / Ghent University (UGent) katrien.strubbe@ugent.be	**Denmark** Niels Bohr Institute, Københavns Universitet / University of Copenhagen (UCPH) Ian Bearden bearden@nbi.ku.dk
Finland Inkeri Kontro Helsingin yliopisto / University of Helsinki (UH) inkeri.kontro@helsinki.fi	**Germany** David Buschhueter Universität Potsdam / University of Potsdam David.buschhueter@uni-potsdam.de
Italy Marta Carli Università degli Studi di Padova / University of Padova (UNIPD) marta.carli.1@unipd.it	**Spain** Diego P. Ruiz Fernando Cornet Universidad de Granada / University of Granada (UGR) druiz@ugr.es
The Netherlands Hay Geurts Radboud Universiteit Nijmegen / Radboud University (RU) hay.geurts@science.ru.nl	

1. Introduction to the subject area of Physics

1.1 Overview

Physics is concerned with the quantitative description of both natural and human-made systems and phenomena on the basis of reproducible experimental findings and rigorous mathematical tools. It is one of the most fundamental disciplines, the knowledge base of which is continuously advancing, and it deals with the most profound questions about our Universe. Physics provides a conceptual basis for other disciplines, and it is involved with the cultural, societal, and technological issues of our time.

Physics is both a theoretical and an experimental subject. Direct observation and measurements are an integral part of physics, and agreement with experiments constitutes the final validity test for any physics theory. Theories, on the other hand, are based on models described by sophisticated mathematics equations. Competence in mathematical modeling and acquaintance with mathematical methods is necessary for solving physics problems. In fact, mathematics plays a structural role in physics, and the development of the two disciplines is strongly intertwined.

Physics is also connected with technology: many physics discoveries have led to technological revolutions, and technological innovations have enabled advancements in physics. Moreover, today computers and IT technologies play an inestimable role in the construction of physics theories and in physics experiments.

1.2. Education and professional context

Due to their wide and deep knowledge base, their solid analytical, computational, and practical skills, and their flexible competences, physics graduates are eligible for a great variety of jobs, from research careers to industry, and even in areas such as business and finance. Teaching in secondary schools is also an important option in many countries. Last but not least, many physics graduates continue their studies in the third cycle.

To prepare such graduates, degree courses in physics must provide a strong fundamental base in physics and mathematics, which is then complemented by more advanced and specialized subjects. Programs need to be flexible enough to enable both research- and job-oriented careers, and to incorporate advancements of the discipline.

An important aspect of the physics community is its international character, in both research and education. This suggests the need and the opportunity for physics students to be mobile during their learning path. For this reason, mobility and exchange programmes between European countries play a major role in physics degree courses.

1.3. Physics degree programmes in Europe

In Europe, most physics degree programs are based on:

1) A bachelor's degree with a common core in classical physics and elements of modern physics.
2) A master's degree focused on a specialistic branch of physics (e.g. theoretical physics, astrophysics, condensed matter physics, etc.) or an interdisciplinary masters' degree. Many of these programmes are research oriented.

There are then some exceptions, such as courses where the initial years of the programme are common to different branches of Science, and Physics is chosen later as the majoring subject.
Physics degrees are usually offered within the faculty of Science. In other cases, the offer of a degree course in Applied Physics or similar occurs within the faculty of Engineering.
Besides physics degree programmes, Physics departments often deliver degree courses in physics-related subjects, both research- and job-oriented, sometimes in collaboration with other scientific departments. Examples of such courses are Astronomy, Optics and Optometry, Materials Science.

1.4 The CAHLOHEE Qualifications and Assessment Reference Frameworks for Physics

Moving from this analysis, in the CALOHEE project (phase 1) the Physics SAG developed a Physics Qualifications Reference Framework (QRF) and a related Assessment Reference Framework (ARF) which is summarized in Figure 1. Nine subject-specific dimensions and different sub-dimensions containing the actual measurable learning outcomes were identified.

PHYSICS

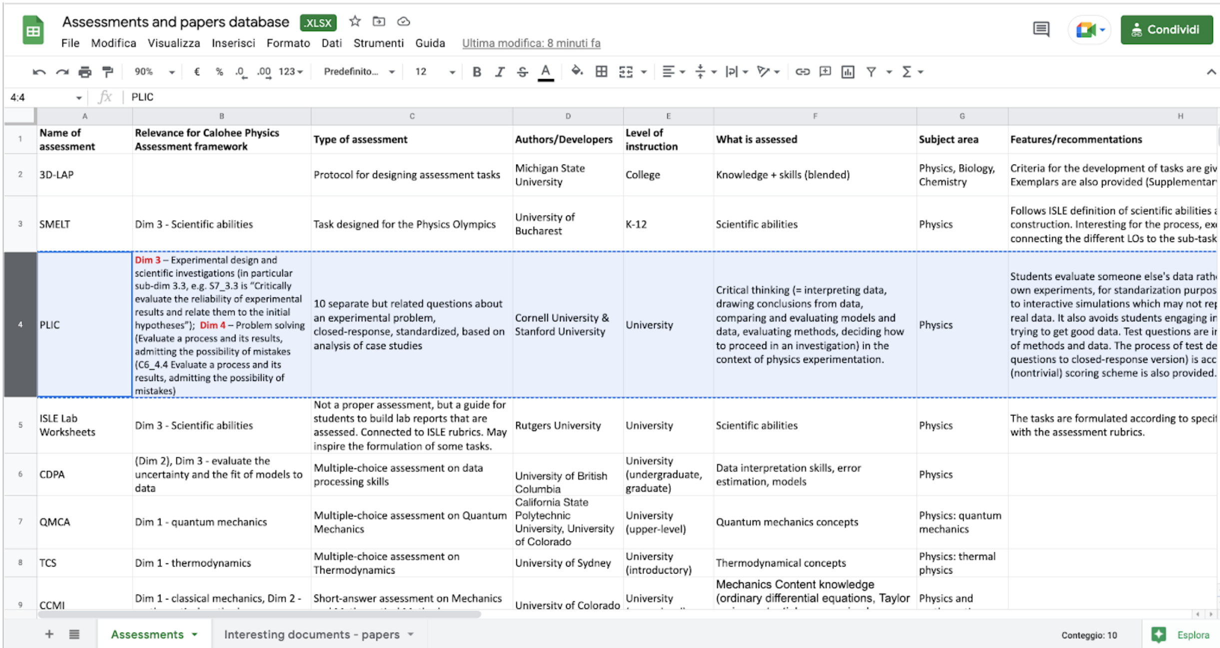
Dimensions of the QF-EHEA	Dimensions of the CALOHEE Framework	Sub-dimensions
I) Knowledge and Understanding	1. Theories and models	1.1 Theories and phenomena 1.2 Applications of theories and models
	2. Mathematical methods	2.1 Mathematical tools 2.2. Computational tools
II) Applying Knowledge and Understanding	3. Experimental design and scientific inquiry	3.1 Experimental design and methodology 3.2 Instrumentation 3.3 Data analysis 3.4 Experiment documentation 3.5 Safety
	4. Problem solving	4.1 Problem framing 4.2 Analytical thinking 4.3 Solution procedure and execution 4.4 Validation of results 4.5 Creative and innovative thinking
III) Making Judgement	5. Scientific culture	5.1 History of Physics 5.2 Epistemology 5.3 Sources of scientific information
	6. Work ethic and integrity	6.1 Ethical rules in the profession 6.2 Awareness of professional actions impact 6.3 Governance and decision making
IV) Communication Skills	7. Communication	7.1 Information sources 7.2 Data representation 7.3 Means of communication 7.4 Technical English
	8. Project management and teamwork	8.1 Project management tools 8.2 Teamwork 8.3 Organizations, societies and cultures
V) Learning skills	9. Professional development	9.1 Professional requirements 9.2 Personal capacities and attitudes

Figure 1 The CALOHEE Assessment Reference Framework for Physics.

2. Assessments – Definitions and Options

2.1. Models of assessment and testing in physics

In order to identify the assessment models that best apply to the Physics subject area, the SAG collected examples of tasks, approaches, and criteria that have been used for physics assessment and testing and analyzed them in terms of their applicability in an international context. The examples came from both the literature and the SAG members' own experience. As we found and analyzed the assessments, we compiled an "Assessments and papers database", an excerpt of which is reported in Figure 2.


Assessments and papers database .XLSX
File Modifica Visualizza Inserisci Formato Dati Strumenti Guida Ultima modifica: 8 minuti fa
Condividi
4:4 PLIC

	A	B	C	D	E	F	G	H
1	Name of assessment	Relevance for Calohee Physics Assessment framework	Type of assessment	Authors/Developers	Level of instruction	What is assessed	Subject area	Features/recommentations
2	3D-LAP		Protocol for designing assessment tasks	Michigan State University	College	Knowledge + skills (blended)	Physics, Biology, Chemistry	Criteria for the development of tasks are giv Exemplars are also provided (Supplementar
3	SMELT	Dim 3 - Scientific abilities	Task designed for the Physics Olympics	University of Bucharest	K-12	Scientific abilities	Physics	Follows ISLE definition of scientific abilities a construction. Interesting for the process, ex connecting the different LOs to the sub-task
4	PLIC	Dim 3 – Experimental design and scientific investigations (in particular sub-dim 3.3, e.g. S7_3.3 is "Critically evaluate the reliability of experimental results and relate them to the initial hypotheses"); Dim 4 – Problem solving (Evaluate a process and its results, admitting the possibility of mistakes (C6_4.4 Evaluate a process and its results, admitting the possibility of mistakes)	10 separate but related questions about an experimental problem, closed-response, standardized, based on analysis of case studies	Cornell University & Stanford University	University	Critical thinking (= interpreting data, drawing conclusions from data, comparing and evaluating models and data, evaluating methods, deciding how to proceed in an investigation) in the context of physics experimentation.	Physics	Students evaluate someone else's data rath own experiments, for standarization purpos to interactive simulations which may not rep real data. It also avoids students engaging ir trying to get good data. Test questions are ir of methods and data. The process of test de questions to closed-response version) is acc (nontrivial) scoring scheme is also provided.
5	ISLE Lab Worksheets	Dim 3 - Scientific abilities	Not a proper assessment, but a guide for students to build lab reports that are assessed. Connected to ISLE rubrics. May inspire the formulation of some tasks.	Rutgers University	University	Scientific abilities	Physics	The tasks are formulated according to specif with the assessment rubrics.
6	CDPA	(Dim 2), Dim 3 - evaluate the uncertainty and the fit of models to data	Multiple-choice assessment on data processing skills	University of British Columbia	University (undergraduate, graduate)	Data interpretation skills, error estimation, models	Physics	
7	QMCA	Dim 1 - quantum mechanics	Multiple-choice assessment on Quantum Mechanics	California State Polytechnic University, University of Colorado	University (upper-level)	Quantum mechanics concepts	Physics: quantum mechanics	
8	TCS	Dim 1 - thermodynamics	Multiple-choice assessment on Thermodynamics	University of Sydney	University (introductory)	Thermodynamical concepts	Physics: thermal physics	
9	CCMI	Dim 1 - classical mechanics, Dim 2 -	Short-answer assessment on Mechanics	University of Colorado	University	Mechanics Content knowledge (ordinary differential equations, Taylor	Physics and	

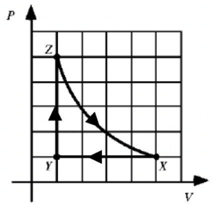
Assessments | Interesting documents - papers | Conteggio: 10 | Esplora

Figure 2 An excerpt of the "Assessments and papers database" compiled by the Physics SAG.

From the analysis we identified different families of assessments:

Concept inventories and multiple-choice tests (e.g., Wattanakasiwich et al., 2013; many more on https://www.physport.org). They assess sophisticated knowledge and sometimes skills, and they are mostly referred to Dimensions 1 and 2 of the CALOHEE Physics Qualifications/Assessment Reference Framework. Although their format is suitable for large-scale assessment, they are not meant to assess the "autonomy and responsibility" level. In Figure 3 we report an example of such an inventory (Conceptual Survey in Thermodynamics, Wattanakasiwich et al., 2013).

Please use the following information to answer questions 29-31.
A student performs an experiment with an ideal gas that is contained in a cylinder with a piston. The P-V diagram below shows the values of pressure and volume of the gas throughout the experiment, starting at point X, continuing to points Y and Z, and returning to point X. Process $Z \rightarrow X$ is isothermal.



29. What is the total work done by the gas in the entire cycle ($X \rightarrow Y \rightarrow Z \rightarrow X$)?
A) Positive B) Negative C) Zero

30. What is the total heat transfer for the entire cycle ($X \rightarrow Y \rightarrow Z \rightarrow X$)?
A) Positive B) Negative C) Zero

31. What is the change of internal energy of the gas in the entire cycle ($X \rightarrow Y \rightarrow Z \rightarrow X$)?
A) Positive B) Negative C) Zero

Figure 3. A set of questions from the Conceptual Survey in Thermodynamics, Wattanakasiwich et al., 2013.

Tasks for evaluating "scientific abilities". These tasks operate at the skills or wider competence level and are usually intended for evaluating students' work in a real setting with manual evaluation from the instructor. The assessment in this case is often carried out using assessment rubrics (e.g., Etkina, 2006). In Figure 4 we report an example of laboratory task from the ISLE model developed by Etkina's group at Rutgers University and an excerpt of a related assessment rubric.

Available equipment: Optics bench, 2 converging (convex) lenses with different focal lengths, meter stick, light source w. crosshairs pattern (the object), viewing screen.

Note: The light source can produce a set of parallel beams of white light as well if you need it to.

For the experiment, write the following in your report:

a) First, recall the thin lens equation. Describe what quantities it relates. What assumptions are inherent in the equation?
b) Design two experiments to test the hypothesis. The two experiments should involve distinctly different situations. Describe your procedure for each. Include ray diagrams.
c) What outcome does the hypothesis predict for each experiment?
d) Conduct the experiments and record the data. Include uncertainties.
e) Were the outcomes consistent with the predictions?
f) What is your judgment about the hypothesis you are testing?

(a)

RUBRIC C: Ability to design and conduct a testing experiment (testing an idea/hypothesis/explanation or mathematical relation)					
Scientific Ability		Missing	Inadequate	Needs some improvement	Adequate
C4	Is able to make a reasonable prediction based on a hypothesis	No attempt to make a prediction is made.	A prediction is made that is distinct from the hypothesis but is not based on it.	A prediction is made that follows from the hypothesis but there are reasoning errors	A correct prediction is made that follows from the hypothesis.
C7	Is able to decide whether the prediction and the outcome agree/disagree	No mention of whether the prediction and outcome agree/disagree.	A decision about the agreement/disagreement is made but is not consistent with the outcome of the experiment.	A reasonable decision about the agreement/disagreement is made but experimental uncertainty is not taken into account.	A reasonable decision about the agreement/disagreement is made and experimental uncertainty is taken into account.

(b)

Figure 4. An example of (a) laboratory task and (b) assessment rubric from the ISLE model (see Etkina et al., 2006).

Another example of task (from Laverty et al., 2016; Michigan State University) is reported in Figure 5. Here, the authors highlight how such a task assesses knowledge and skills in a blended way.

Physics Example 3: Design an experiment, including the measurements needed, to test the claim that the current is the same everywhere in a series circuit. Explain the rationale for your design, including any underlying assumptions you've made.

CRITERIA	RATIONALE
Scientific Practice: Planning Investigations	
1. Question poses a scientific question, claim, or hypothesis to be investigated. 2. Question asks student to describe or design an investigation, or identify the observations required to answer the question, to answer the question or test the claim or hypothesis. 3. Question asks student to justify how their description, design, or observations can be used to answer the question or test the claim or hypothesis.	1. Question makes the claim that "the current is the same everywhere in a series circuit". 2. Questions asks student to "Design an experiment". 3. Question asks student to "Explain the rationale for your design."
Crosscutting Concept: Not Applicable	
Core Idea: Interactions are Mediated by Fields	
Fields are generated by charges/masses. Fields affect charges/masses. In circuits, fields induce currents.	Question asks student to identify the measurements needed to evaluate the current everywhere in a circuit.

Figure 5. An example of task assessing knowledge and skills in a blended way (from Laverty et al., 2016).

Tasks for evaluating problem solving. Similar to the previous case, these tasks operate at the skills or wider competence level, are intended for evaluating written student work in a real setting with manual evaluation, and involve the use of assessment rubrics. In Figure 6 we report an example from Docktor et al., 2016.

You are designing part of a machine to detect carbon monoxide (CO) molecules (28 g/mol) in a sample of air. In this part, ultraviolet light is used to produce singly charged ions (molecules with just one missing electron) from air molecules at one side of a chamber. A uniform electric field then accelerates these ions from rest through a distance of 0.8 m through a hole in the other side of the chamber. Your job is to calculate the direction and magnitude of the electric field needed so that CO^+ ions created at rest at one end will have a speed of 8×10^4 m/s when they exit the other side.

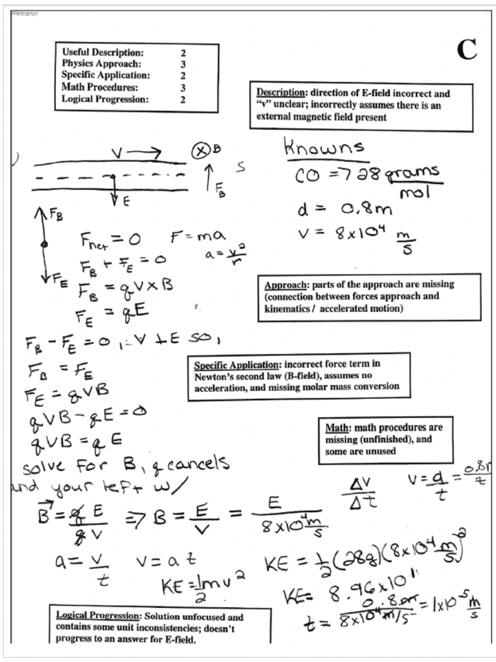
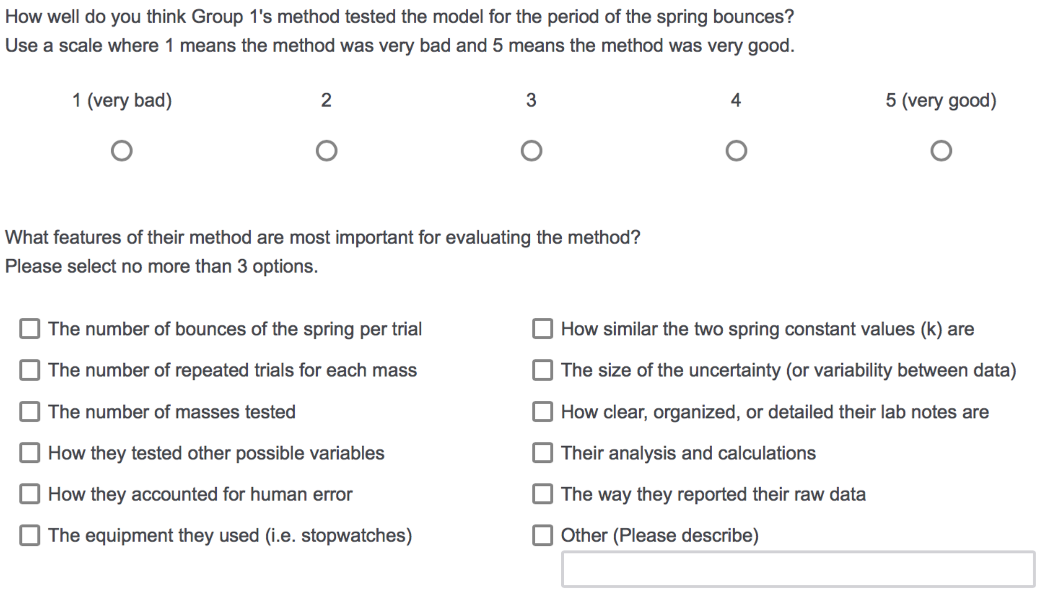


TABLE I.

	5	4	3	2	1	0	NA(problem)	NA(solver)
USEFUL DESCRIPTION	The description is useful, appropriate, and complete.	The description is useful but contains minor omissions or errors.	Parts of the description are not useful, missing, and/or contain errors.	Most of the description is not useful, missing, and/or contains errors.	The entire description is not useful and/or contains errors.	The solution does not include a description and it is necessary for this problem /solver.	A description is not necessary for this *problem*. (i.e., it is given in the problem statement)	A description is not necessary for this *solver*.
PHYSICS APPROACH	The physics approach is appropriate and complete.	The physics approach contains minor omissions or errors.	Some concepts and principles of the physics approach are missing and/or inappropriate.	Most of the physics approach is missing and/or inappropriate.	All of the chosen concepts and principles are inappropriate.	The solution does not indicate an approach, and it is necessary for this problem/ solver.	An explicit physics approach is not necessary for this *problem*. (i.e., it is given in the problem)	An explicit physics approach is not necessary for this *solver*.
SPECIFIC APPLICATION OF PHYSICS	The specific application of physics is appropriate and complete.	The specific application of physics contains minor omissions or errors.	Parts of the specific application of physics are missing and/or contain errors.	Most of the specific application of physics is missing and/or contains errors.	The entire specific application is inappropriate and/or contains errors.	The solution does not indicate an application of physics and it is necessary.	Specific application of physics is not necessary for this *problem*.	Specific application of physics is not necessary for this *solver*.
MATHEMATICAL PROCEDURES	The mathematical procedures are appropriate and complete.	Appropriate mathematical procedures are used with minor omissions or errors.	Parts of the mathematical procedures are missing and/or contain errors.	Most of the mathematical procedures are missing and/or contain errors.	All mathematical procedures are inappropriate and/or contain errors.	There is no evidence of mathematical procedures, and they are necessary.	Mathematical procedures are not necessary for this *problem* or are very simple.	Mathematical procedures are not necessary for this *solver*.
LOGICAL PROGRESSION	The entire problem solution is clear, focused, and logically connected.	The solution is clear and focused with minor inconsistencies	Parts of the solution are unclear, unfocused, and/ or inconsistent.	Most of the solution parts are unclear, unfocused, and/ or inconsistent.	The entire solution is unclear, unfocused, and/ or inconsistent.	There is no evidence of logical progression, and it is necessary.	Logical progression is not necessary for this *problem*. (i.e., one-step)	Logical progression is not necessary for this *solver*.

Figure 6. An example of problem-solving task (a), a student's solution (b) and the rubric used for assessment (c), as proposed by Docktor et al. (2016).

Assessments that try to integrate generic skills with physics core competence skills. These assessments are rarer. An example is the PLIC (Physics Lab Inventory of Critical thinking; Walsh et al., 2019, Cornell University & Stanford University), which was designed to assess critical thinking in the context of physics experimentation. Some skills/competences that are covered in this assessment are interpreting data, evaluating experimental procedures, comparing methods, drawing conclusions from data, and deciding how to proceed in an investigation. The PLIC is thought for an online setting and is an example of how a task related to experimental investigation can be delivered electronically, covering at least some of the subdimensions (Figure 7).



How well do you think Group 1's method tested the model for the period of the spring bounces?
Use a scale where 1 means the method was very bad and 5 means the method was very good.

1 (very bad)	2	3	4	5 (very good)
○	○	○	○	○

What features of their method are most important for evaluating the method?
Please select no more than 3 options.

- ☐ The number of bounces of the spring per trial
- ☐ The number of repeated trials for each mass
- ☐ The number of masses tested
- ☐ How they tested other possible variables
- ☐ How they accounted for human error
- ☐ The equipment they used (i.e. stopwatches)
- ☐ How similar the two spring constant values (k) are
- ☐ The size of the uncertainty (or variability between data)
- ☐ How clear, organized, or detailed their lab notes are
- ☐ Their analysis and calculations
- ☐ The way they reported their raw data
- ☐ Other (Please describe)

Figure 7. An excerpt of the PLIC (Physics Lab Inventory of Critical thinking) assessment, Walsh et al., 2019.

Assessments based on machine learning. These assessments aim at assessing complex tasks in an automated way (e.g., Zhai et al., 2020). The literature on these tasks is however still limited and at a preliminary stage. A further discussion on this type of assessment is reported in Section 4.

2.2 Updated Framework and Topical Issues

The SAG discussion highlighted the points where the existing CALOHEE RF subdimensions and descriptions were suboptimal for the purpose of assessment. Through a feedback process, modifications to the (sub)dimensions were proposed when needed. In particular, the SAG focused their work on Dimension 3 (Experimental design and scientific investigation), Dimension 4 (Problem Solving) and Dimension 7 (Communication) of the CALOHEE Assessment Reference Framework at level 6 (Bachelor), for which a common core among different countries could be more easily found.
Below we report the revised versions of each of the three Dimensions of the Calohee ARF. Here we summarize the main changes:

- For Dimension 3, we split Subdimension 3.1 ("Experimental Design and Methodology") in two subdimensions ("Experimental Design" and "Experimental Method"). They became the new subdimension 3.1 and 3.3, respectively.
- For Dimension 4, subdimension 4.2 ("Analytical Thinking") was renamed "Modeling" to better capture its complexity and the structural link between mathematics and physics that it involves. The names of Subdimension 4.3 ("Execution"), 4.4 ("Evaluation") and 4.5 ("Revision of the process and creative thinking") were also slightly modified.
- For Dimension 7, the subdimension "Data representation" was removed as it was already captured in aspects of Dimensions 3 and 4. Subdimension 7.3 (formerly 3.4) was renamed from "Technical English" to "Language including scientific English" to include assessment of language aspects in the student's own language.

PHYSICS

For all subdimensions, a revision/rephrasing of the descriptors was performed in order to facilitate the alignment between the Assessment Reference Framework and the assessment rubrics that were developed in CALOHE2.

The SAG also considered the work done by the Topical Issues groups. Although the SAG did not feel the need to add specific subdimensions to the existing framework, suggestions from the Topical Group work were incorporated in the rephrasing of descriptors when deemed appropriate.

Dimension 3: Experimental design and scientific investigation (REVISED)			
	Knowledge	*Skills*	*Autonomy and responsibility*
L6_3 **Experimental design and scientific investigation**	**K6_3** **Describe common methods, instrumentation, techniques, theories, and regulations used in experimental physics.**	**S6_3** **Design a simple experimental investigation, using common instrumentation, follow guidelines and apply common methods, techniques and theories for data collection, analysis, and reporting.**	**C6_3** **Set up and carry out simple scientific investigations safely.**
L6_3.1 **Experimental design**	K6_3.1 Identify and/or describe relevant aspects of a scientific investigation. Identify the physical quantities involved in a situation. Describe relevant models or theories.	S6_3.1 Formulate a hypothesis, propose a testable prediction, and devise a plan to test it. Construct appropriate setup to perform the test or tests required.	C6_3.1 Design appropriate experiments. Identify relevant theoretical framework, model, and/or existing data for comparison. Ensure experimental design and procedures will lead to consistent and relevant results.
L6_3.2 **Instrumentation**	K6_3.2 Describe standard instrumentation used in experimental physics, how it is used, and where its use is valid.	S6_3.2 Use instrumentation appropriate for the task at hand. Ensure instrumentation is kept within appropriate range of operation. Formulate a plan for calibration of equipment and perform said calibration if necessary. Operate equipment safely.	C6_3.2 Choose and employ appropriate instrumentation necessary to carry out a given experiment. Identify situations in which it is unwise to use equipment alone or without appropriate supervision. Use equipment responsibly.
L6_3.3 **Experimental method**	K6_3.3 Describe uncertainties expected in measurements made. Identify necessary experimental accuracy necessary. Describe methods to gather reliable data. Plan activities sufficiently to obtain needed data in the time available.	S6_3.3 Carry out experimental activities systematically. Ensure only one experimental parameter is changed at a time. Apply appropriate procedures to gather a sufficient amount of reliable data.	C6_3.3 Participate in discussion of and decisions about experimental techniques chosen, amount of data needed, and reliability of the data collected.
L6_3.4 **Data analysis**	K6_3.4	S6_3.4	C6_3.4

	Describe methods and techniques for the processing of experimental data. Describe relevant statistical measures for the data under consideration.	Organize and analyze data using appropriate tools and techniques. Document analysis techniques used. Estimate uncertainties on data and/or derived results. Evaluate reliability of data and/or derived results.	Process experimental data, indicate uncertainty on data and/or derived results honestly. Store data and results in a transparent and responsible manner. Independently evaluate the reliability of the results.
L6 _3.5 **Experiment documentation**	K6_3.5 Describe standard methods of recording the details of an experimental activity and storing data.	S6_3.5 Keep a record of the details and steps of an experiment, including the acquisition and analysis of data, and communicate these via the appropriate communication channel.	C6_3.5 Identify the appropriate method to report on an investigation and communicate the results.
L6_3.6 **Safety**	K6_3.6 Describe relevant safety issues, equipment, and regulations of a standard physics laboratory.	S6_3.6 Follow the safety regulations of a standard physics laboratory, including using basic individual protection equipment.	C6_3.6 Identify the safety regulations of a physics laboratory and operate accordingly, including the use of appropriate protection equipment.

Dimension 4: Problem solving (REVISED)			
	Knowledge	*Skills*	*Autonomy and responsibility*
L6_4 **Problem solving**	**K6_4** **Link physics concepts and laws with common strategies, procedures, and criteria for framing, representing, solving, and validating the results of a problem.**	**S6_4** **Categorise problems based on physical principles, use different representations to model a problem, apply common procedures to reach a solution and check for its validity.**	**C6_4** **Address problems from the point of view of physics, devise and carry out a plan for reaching a solution and check its validity, devise, and compare different solutions when applicable.**
L6_4.1 **Problem framing**	K6_4.1 Organise knowledge of physics in terms of principles, theories, and conditions for their validity.	S6_4.1 Categorise problems based on physics principles rather than objects and surface features.	C6_4.1 Identify the physics involved in a problem or situation, identify common structures in different problems.
L6_4.2 **Modeling**	K6_4.2 Connect mathematical concepts and ideas to physics concepts.	S6_4.2 Identify the variables involved in a problem and the relationships between them. Use different representations of physics concepts and situations.	C6_4.2 Model the problem mathematically, using different representations (graphs, diagrams, equations) in a consistent way.

	Be familiar with different representations of physics concepts and situations (graphs, diagrams, equations).		
L6_4.3 **Execution**	K6_4.3 Have knowledge of common mathematical methods and procedures used in physics problem solving.	S6_4.3 Apply common mathematical methods and procedures to solve a problem.	C6_4.3 Choose an appropriate and effective mathematical procedure to solve a problem and execute it in a correct and complete way, justifying any assumptions and simplifications.
L6_4.4 **Evaluation**	K6_4.4 Recall common criteria for stating the validity of a result, e.g., evaluating the expected order of magnitude and units.	S6_4.4 Apply common criteria to check the validity of a solution and/or intermediate results.	C6_4.4 Critically evaluate the solution of a problem, discussing its physical meaning and including limiting cases when applicable.
L6_4.5 **Revision of the process and creative thinking**	K6_4.5 Acknowledge the possibility of alternative ways to look at the same problem.	S6_4.5 Revise and evaluate the whole process. Compare own solution with others' solution.	C6_4.5 Recognize ways to improve/extend the validity of a solution. Devise alternative ways to address a problem and critically evaluate them to choose the best one.

Dimension 7: Communication (REVISED)			
	Knowledge	*Skills*	*Autonomy and responsibility*
L6_7 **Communication**	**K6_7** **Identify reliable sources of scientific information, recognize the specificities of different means used to communicate science, and have sufficient language knowledge to communicate science topics both in own language and in English.**	**S6_7** **Search for scientific information and evaluate its reliability, present scientific information with appropriate language (including the English language) orally and in writing.**	**C6_7** **Retrieve and appropriately use scientific information also to support arguments; communicate scientific results using a correct and appropriate language in different situations and be able to study scientific topics in English.**
L6_7.1 **Information sources**	K6_7.1 Identify the criteria that make an information source reliable.	S6_7.1 Evaluate the reliability of a source of scientific information.	C6_7.1 Retrieve and use appropriate and reliable sources of information about a scientific topic.

	Have knowledge of the main tools and methods used to retrieve scientific information sources (e.g., knowledge of libraries and article databases).	Search for scientific information at the desired level using appropriate tools.	Support an argument using reliable sources, also concerning societal issues.
L6_7.2 **Means of communication**	K6_7.2 Recognize the specificities of different means of communication (scientific journals, books, newspapers, lab reports, oral/written presentations, videos, etc.) with respect to science communication and dissemination of scientific results.	S6_7.2 Produce scientific reports and presentations, oral or written, respecting the given criteria (length of presentation, etc.).	C6_7.2 Identify and choose the appropriate written or oral style according to the context, audience, and communicative goals (laboratory report, dissertation, scientific article, etc.).
L6_7.3 **Language including scientific English**	K6_7.3 Demonstrate knowledge of technical and scientific terms in the language(s) used in the country. Demonstrate knowledge of the English language at the level necessary for understanding general physics texts and communicating simple physics topics in English (minimum B1 level).	S6_7.3 Produce a well-structured, concise, and linguistically correct report or presentation. Read, speak, write in scientific English (minimum B1 level).	C6_7.3 Communicate physics topics using appropriate terms both in own language and in English. Autonomously study physics topics in English, e.g., understanding and reporting about a scientific paper.

3. Exploration Process

In order to identify the dimensions to be addressed in a comparative assessment, the SAG started by analyzing existing degree programmes in physics and matching them against the CALOHEE frameworks. A dedicated document has been delivered in this regard and we refer to it for a detailed analysis of the issue. Here we report a synthesis of the SAG conclusions useful for understanding and contextualizing the SAG work on the assessment.
The analysis of programmes from six different countries (Belgium, Finland, Germany, Italy, Spain, The Netherlands) revealed that the emphasis of most programmes is on the core disciplinary areas (Dimensions 1 and 2 of the CALOHEE QRF/ARF), experimental design (Dimension 3) and problem solving (Dimension 4).
This situation reflects some of the specificities of the physics subject outlined above: training a graduate able to work at the frontiers of physics requires an advanced knowledge and skills base that takes far beyond the first years of the bachelor programme to be developed, and many physics programmes therefore devote a large part of their courses to these fundamental competences.
Among the other dimensions, communication (Dimension 7) is the best developed, while for the others the situation is variable and often limited to some sub-dimensions. In almost all cases, these sub-dimensions are integrated in the core disciplinary dimensions, rather than being developed as standalone competences.

3.1 SAG Discussions

3.1.1 Challenge and opportunities

Reflecting on the subject specificities outlined above, the SAG identified some opportunities and challenges that characterize the physics subject area with respect to the purpose of developing an international comparative assessment.

Opportunities:

- The physics knowledge base is independent on the specific country, especially at the bachelor level. Common cores in classical and modern physics can be identified.
- There is a wealth of results from physics education research on the assessment of students' understanding of specific physics topics. Physics Education Research (PER) is a lively research field with a rich tradition both in Europe and outside. It is concerned with the teaching and learning of physics at all levels, and it has produced many established results (including assessments) that are available for the physics community, for example through the PhysPort platform (https://www.physport.org/).

Challenges:

- At the master's degree level, students specialize in specific branches of physics that can significantly differ from one another in terms of knowledge base and methods. It is more difficult to find a common ground for advanced levels.

- Experimentation is a major part of physics, but it is difficult to assess through large-scale tests. In fact, experimental competences are ideally assessed in a real-lab setting.
- A similar challenge applies to problem solving. An authentic problem-solving addressing the "autonomy and responsibility" level cannot be assessed using closed-ended questions, which are more suitable for the knowledge or skills level.
- Unlike subject-specific competences, generic competences are often not explicitly mentioned in physics degree programs and is therefore less clear how to assess them.

3.1.2 Reinterpretation of the CALOHEE Framework: interplay of dimensions

To address these challenges, the Physics SAG reflected on the CALOHEE QRF and analyzed the relationship between the different dimensions. We recognized that the 9 dimensions play different roles in the development of the physics graduate profile. In particular, we agreed that for the purpose of designing an assessment the Physics QRF should be reinterpreted as summarized in Figure 8 (left).

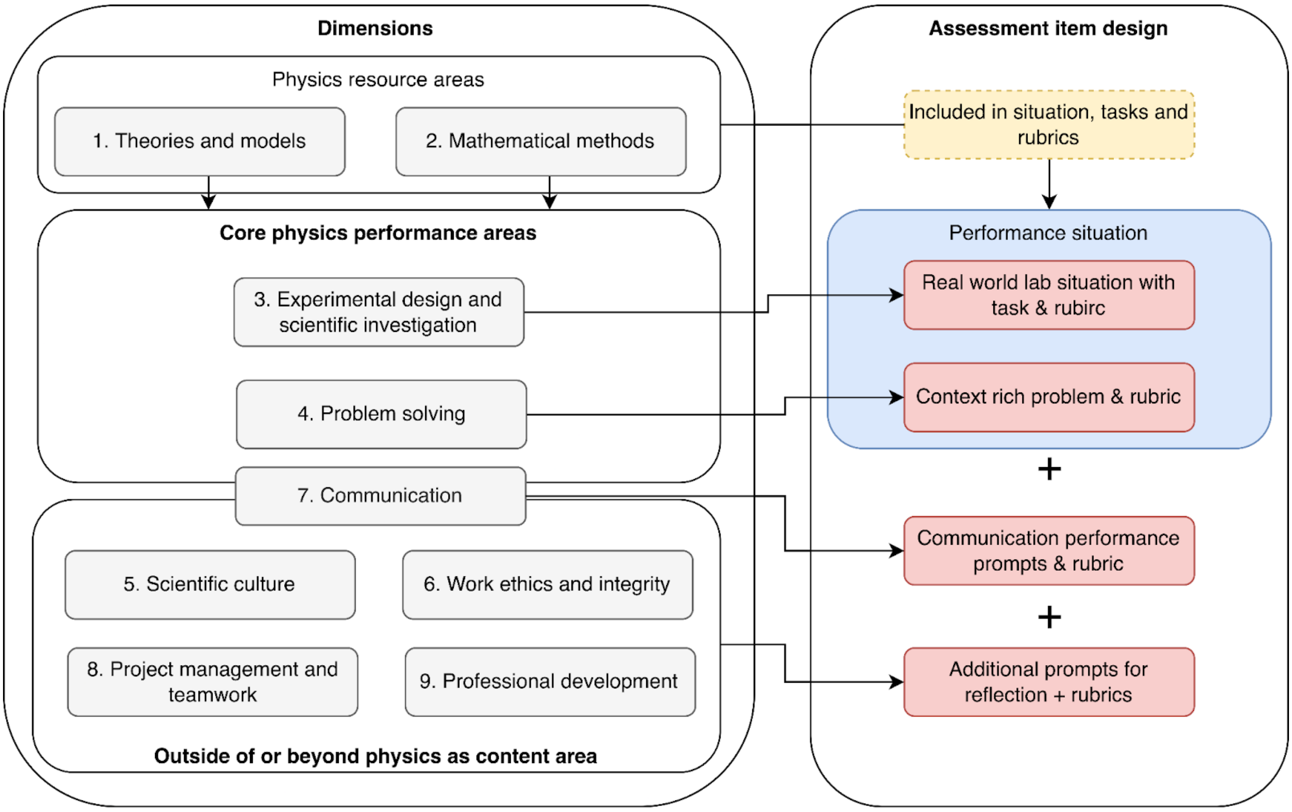


Figure 8. (Left) Reinterpretation of the Physics CALOHEE QRF and (right) its connection with item design

Dimensions 1 (Theories and models) and 2 (Mathematical methods) constitute what we called the "Physics resource areas". They contain the knowledge and skills base required to physics graduates, that constitute the cognitive resources for competence in the other dimensions. The levels of knowledge and skills of dimensions 1 and 2 can be easily assessed using common end-of-chapter exercises or existing test banks. However, it is harder to identify an assessment at the "autonomy and responsibility" level that is clearly distinguished from tasks related to problem solving or experimental design. In practice, dimensions 1 and 2 at the autonomy and responsibility level are included, in practical situations, in tasks and rubrics

related to dimensions 3 (Experimental design and scientific investigation) and 4 (Problem solving).
Dimensions 3 and 4 are, in fact, the "Core physics performance areas"[1], those where a physicist actually orchestrates her/his resources in order to solve a complex task. They are the ones in which, to our opinion, actual subject-specific *competence* is involved, and the ones where an assessment at the "autonomy and responsibility" level should be focused.
The remaining dimensions (generic competences) should be integrated in the physics graduate profile, without however constituting the core performance areas of physics. Among them, communication (dimension 7) occupies a special position, being strictly connected with both experimental competence and problem solving. We believe that the assessment of these competences should be integrated in the physics core performance areas, rather than treated as standalone. Task-specific prompts can be constructed to this end.
Based on these reflections, the SAG decided to focus the CALOHE2 assessment blueprint on three dimensions:

- Dimension 3: Experimental design and scientific investigation.
- Dimension 4: Problem solving.
- Dimension 7: Communication.

3.1.3 Assessments: Purposes, Formats, and Rubrics

The SAG discussed for long whether the proposed items should be designed to be directly applied for large-scale testing. We considered different formats and for each one we discussed the pros and cons. However, the simultaneous identification of an "ideal" assessment related to the CALOHEE RF and its transformation for large-scale testing seemed to us a too big step at this stage. We therefore decided to focus on the development of assessment items, leaving considerations about how the assessment could be transformed for large-scale testing as a second step. We also decided to focus on level 6, since a common ground about the topics to be included could be more easily found. We do not see this as a limitation as the same logic presented here can be applied to more sophisticated topics or experimental situations according to the physics specialization to be tested.
The assessment logic and structure are represented in the right part of Figure 8. For our "core performance" dimensions (3 and 4) we decided to develop authentic "performance situations" consisting in a real-world lab task and a context-rich problem, respectively. Dimension 7 is then integrated through two sub-tasks connected with each authentic performance situation.

In **Dimension 3**, we want to assess whether students can set up and carry out simple scientific investigations safely. This is evaluated by assigning a real-world situation where students have to design a reliable experiment, use the available instrumentation in a safe way to collect and analyze the data, document the experiment, and compare the results with physics theories and models or with data from other sources.
The discussion whether such a situation is assessing the "skills" level, or the desired "autonomy and responsibility" level was not trivial. We agreed that what distinguishes the competence level from the skills level is the degree of *autonomy* students have in, for

[1] To the expert reader, this division may remind of the QF-EHEA classification, from which, in fact, our RF originates. Our proposed sub-division makes it more subject-specific and assessment-oriented.

example, choosing the equipment and the experimental techniques. A task at the skills level would, for example, include scaffolding for each step and only the required instrumentation would be provided to the students. For a task at the competence level the scaffolding would be removed, and complexity can be added by, for example, including equipment that is not needed for the experiment and/or providing alternative equipment so that the students have to choose what to use for their experiment and to justify their choices. We also commented that skills could be evaluated by proposing specific, independent tasks for each of the steps involved in experimentation, while an assessment of competence requires a complex task where the different stages of experimentation are connected.

To assess students' performance in this type of tasks, we developed a rubric (the "Experimental design and investigation rubric"), inspired by the literature (in particular, Etkina et al., 2006), in which each indicator is matched to one of the (revised) sub-dimensions of Dimension 3. In the rubric, some sub-dimensions are further split into smaller items to allow flexibility in the use of the rubric for different experimental situations.

The type of problems that apply to **Dimension 4** cannot be solved by just recalling knowledge. They should, nevertheless, imply the knowledge of physics theories and models, and require modeling skills, sometimes called "mathematization" skills in the literature.

Also for Dimension 4 we discussed what would distinguish the skills and competence level. In this case, we agreed that what distinguishes them is the level of *authenticity* of the problem. In fact, in a problem addressing the "skills" column you would "just" look for the ability of going through the different steps of problem solving and may not want to distract the students with irrelevant elements; on the contrary, when evaluating autonomy and responsibility you look for these abilities transferred in an authentic situation. Students should take decisions about, for example, what parameters should be considered and how they should be interpreted in the context of the problem, which should reflect a real-life, rather than idealized, situation. Context-rich and sometimes even ill-defined problems, where students have to extract data or info from documents provided with the task or estimate some data, are therefore the type of problems we are looking for.

In order to assess this type of problems we constructed a rubric (the "Problem Solving Rubric"), matched to the subdimensions of the ARF and applicable to different problem situations. Also this rubric was inspired by works in the literature (e.g., Docktor et al., 2016).

Finally, for **Dimension 7** we did not construct independent assessments, but we thought of prompts to be added to the assessment task developed for Dimensions 3 and 4. We also built an assessment rubric, the "Communication in Physics Rubric", to evaluate students' work on these prompts.

4. Actual Assessments

In this section we present assessment rubrics and examples of assessment tasks for the chosen dimensions. More specifically, this section is organized as follows. For each of the two core performance dimensions (Dimension3 – Experimental design and scientific investigation; Dimension 4 – Problem Solving), we provide:

- The assessment rubric;
- Two examples of assessment tasks and an analysis of each task using the rubric;
- A possible related prompt/task for assessing Dimension 7 (Communication).

Finally, we provide the rubric for Dimension 7.

4.1. Rubrics and examples of assessment tasks

Dimension 3

In the next pages, we report the “Experimental design and investigation rubric” developed for assessing Dimension 3. The matching with the revised CALOHEE ARF subdimensions is indicated.

EXPERIMENTAL DESIGN AND INVESTIGATION RUBRIC					
INDICATOR	**0**	**1**	**2**	**3**	**Not applicable**
L6_3.1 **Experimental design**	C6_3.1 Design appropriate experiments. Identify relevant theoretical framework, model, and/or existing data for comparison. Ensure experimental design and procedures will lead to consistent and relevant results.				
3.1.1 **Understand the problem at hand**	Unable to translate question to action.	State what needs to be measured.	State what needs to be measured, realize which tools are best suited. Sketches situation.	Grasps what needs to be measured and why it is interesting.	
3.1.2 **Identify relevant variables**	No or wrong variables identified.	Dependent and independent variables identified, but not only those that are relevant and/or necessary.	Correct variables identified.		
3.1.3 **Identify appropriate theory/model**	No attempt made to connect experiment with theory or model.	There are references to connected theories/models but include mistakes.	The appropriate theory/model is identified.		This measurement requires no model or theory for comparison.
3.1.4 **Make a testable hypothesis**	Unable to make hypothesis or hypothesis is not testable.	Hypothesis made and testable in principle.	Hypothesis made and testable in time given with equipment at hand.		Experiment not amenable to hypothesis testing.
3.1.5 **Design reliable experiment**	Experimental design insufficient for task at hand.	Design works but leads to inconsistent results.	Design is robust and reliable but requires substantial optimization.	Design robust and reliable, requires no further optimization.	
L6_3.2 **Instrumentation**	C6_3.2 Choose and employ appropriate instrumentation necessary to carry out a given experiment. Identify situations in which it is unwise to use equipment alone or without appropriate supervision. Use equipment responsibly.				
3.2.1 **Identify appropriate equipment**	Inappropriate equipment identified (e.g., instruments missing and/or includes elements that are not needed).	Appropriate equipment chosen, but not optimal or includes extraneous elements.	Optimal equipment chosen for the experiment to be performed.	Optimal equipment chosen and reasons for this choice are given.	

3.2.2 **Construct/assemble experiment**	Final experiment is poorly constructed (mechanical and/or electrical connection don't work; tape is used; experiment unsafe by construction).	Experiment setup works and does not present significant issues.	Experimental setup works; setup cleanly constructed and correctly characterized (e.g., length of pendulum correctly measured).	Experimental setup is optimal and well characterized.	Experimental apparatus provided to student.
3.2.3 **Optimization**	No attempt to optimize setup.	Attempts to optimize setup to minimize uncertainty.	Some optimization to the setup is made.	The setup is fully optimized for minimization of uncertainty (e.g., length of pendulum).	Experiment adequate without further optimization.
L6_3.3 **Experimental method**	C6_3.3 Participate in discussion of and decisions about experimental techniques chosen, amount of data needed, and reliability of the data collected.				
3.3.1 **Conduct experiment(s)**	Changes are not systematic; More than one parameter changed simultaneously.	Attempts to change one variable at a time, repeatability checked for one or two settings.	Variable control is maintained; reliability of experiment is checked thoroughly.		
3.3.2 **Collect data**	Unable to collect data; only one measurement per setting OR several repetitions that are inconsistent.	Multiple measurements per setting, but too few settings OR many settings with insufficient statistics.	Sufficient number of settings, sufficient data at each setting.	Data from a broad range of settings, with higher granularity in region of interest.	
L6_3.4 **Data analysis**	C6_3.4 Process experimental data, indicate uncertainty on data and/or derived results honestly. Store data and results in a transparent and responsible manner. Independently evaluate the reliability of the results.				
3.4.1 **Process data**	Raw data used, no or incorrect uncertainties.	Calculate appropriate means and associated uncertainties.	All uncertainties correctly calculated, attention paid to significant digits.	Uses uncertainties to guide data collection; explicit discussion of collection time vs. statistical uncertainty.	
3.4.2 **Present data**	Data and/or uncertainties not presented or presented incorrectly; plots/tables not	Data presented in appropriate way (tabular or plot); uncertainties included;	Data presented clearly and succinctly; data points and uncertainties clearly delineated; choice of tables.	Data presented very clearly and effectively.	

	clearly labeled; units not included.	plot/table clearly labeled; units included.			
3.4.3 **Comparisons with model, theory, and/or existing data**	No attempt at comparing results with appropriate data/model/theory, OR Data compared incorrectly (quantities not the same; model/theory inapplicable to situation).	Appropriate comparisons made; unable to convincingly reject/accept hypothesis OR Incorrect conclusion drawn from presented data.	Comparisons made, hypothesis rejected or accepted.	Differences and/or ratios of data and comparator are presented, comparison includes discussion of uncertainties in data/model used for comparison, discussion of where comparison is (or is not) valid. Clear and correct conclusions drawn from the data presented.	Activity only involves collecting and understanding data from this experiment.
L6 _3.5 **Experiment documentation**	C6_3.5 Identify the appropriate method to report on an investigation and communicate the results.				
3.5.1 **Documentation**	No log of activity kept.	Notes taken on some aspects of activity.	Sufficient documentation produced that activity could be repeated.	Activities documented, appropriate diagrams, schematics, pictures included.	
L6_3.6 **Safety**	C6_3.6 Identify the safety regulations of a physics laboratory and operate accordingly, including the use of appropriate protection equipment.				
3.6.1 **Safety**	No attention given to safety issues.	Most obvious safety issues noticed.	Experiment presents no dangers to experimenters or others.	Experiment presents no safety issues, and they are explicitly considered and discussed.	

Authentic performance tasks for Dimension 3

The Pendulum task

The physics topic is one all the students are familiar with, as it is part of the common core knowledge base characterizing all bachelor physics courses.

A possible prompt could be:

Given the equipment at your disposal, build a pendulum that allows you to determine the angle at which the small angle approximation is no longer valid.

The task is to be accomplished in a real lab setting equipped with standard tools, including things students don't need (set of masses; different kinds of strings; stopwatch and photogates; tape measure + protractor; scale; power supplies; video camera; sonic/laser rangefinder, etc.). A time span of one hour should be sufficient to complete the task; additional time would be required to produce a report about the experiment.

Analysis of the Pendulum task according to the Experimental design and investigation rubric:

Rubric item(s)	**Expected performance**
Understand the problem at hand Identify relevant variables Identify appropriate theory/model ***L6_3.1***	Calculate (or know/be able to justify) the expected period of a simple pendulum and apply Newton's 2nd law. Derive equation of motion.
Make a testable hypothesis Design reliable experiment ***L6_3.1***	Understand that small angle approximation means sin(angle)≈angle; realize that the small angle approximation yields a simple differential equation they know how to solve ("Hooke's law"); realize that this implies a period that is independent of angle.
Identify appropriate equipment ***L6_3.2***	Choose how to measure time (One period or many? Use stopwatch or high-resolution timer? Where in the pendulum's path is the best place to start/stop timer? How many independent measurements will be required? What is the time resolution of available methods?).
Construct/assemble experiment Optimization ***L6_3.2***	Build a simple pendulum using a light string and heavy mass; hang from two points, realize that length is from center of mass of pendulum bob to the midpoint of the two hanging points; choose smallest bob possible to minimize energy loss to air resistance; realize that a longer pendulum gives a longer period.
Conduct experiment(s) Collect data ***L6_3.3***	Measure period and correct length of pendulum, estimate uncertainties associated with these quantities; take data at an appropriate number of angles; use an appropriate strategy to find the interval of angles of interest; take sufficient time measurements to have a reasonable determination of uncertainties; optimize uncertainty and time spent measuring.
Process data ***L6_3.4***	Calculate mean, std dev, and error on mean for measurements at each angle; plot the data thus obtained appropriately.
Present data Comparisons with model, theory, and/or existing data ***L6_3.4***	Plot data and model together; correctly display uncertainties associated with data points; use comparison to modify experiment or take new data; justify conclusions based on objective criteria (e.g., where data are Nσ from model).
Documentation ***L6_3.5***	Document experiment and results in a logbook; produce a report.

Safety ***L6_3.6***	Operate equipment safely, e.g., do not hang the pendulum in a dangerous or unstable position.

The Speed of Light task

The second proposed task involves a more advanced physics topic (the speed of light), less standard equipment, and more sophisticated experimental techniques, but it still accessible to bachelor students. It is based on an experiment used in one of the SAG members' departments. The following description applies to the specific setup and equipment used in that case, but a similar experiment can be set up in any physics department with rather inexpensive equipment.
Figure 4 shows a scheme of the setup. A blinking LED is used as the light source. Two forward detectors are mounted facing each other, 90° with respect to the source. A beam splitter is placed in front of the LED board to reflect some of the light onto one detector, while the rest is transmitted and focused on a movable reflector using one or two Fresnel lenses. Light is reflected back onto a second detector and the delay between the two pulses is measured by means of an oscilloscope.

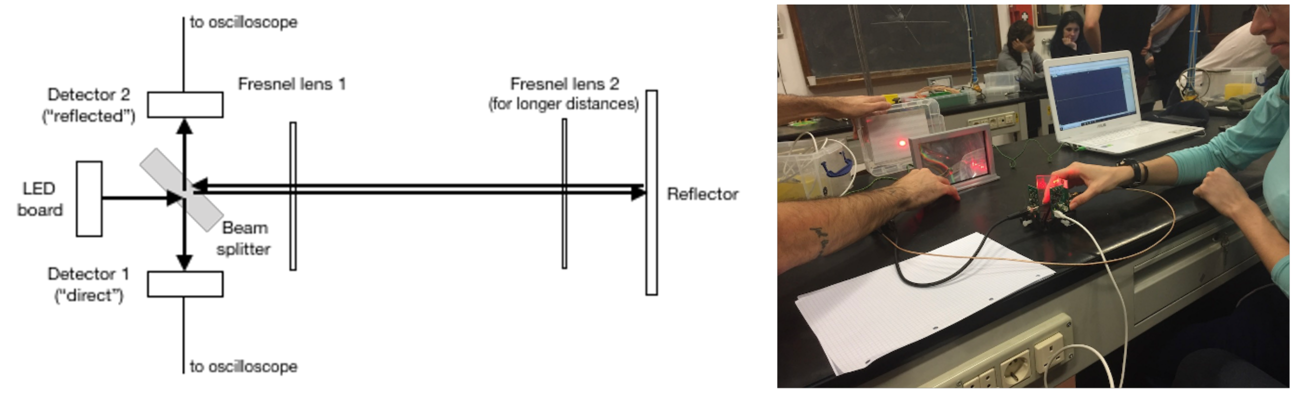


Figure 9. Scheme of the experimental setup for the "Speed of Light" task and students working on the apparatus.

A possible prompt could be:
Given the equipment at your disposal, estimate the speed of light.

In this case, the main setup elements are given, but students still need to set up the actual experiment and have the freedom to select some of the experimental parameters such as the pulse width and the period of the blinking LED, they can choose how to make the measurements (how to measure distances, where on the signal to measure the distance between the two pulses, etc.), they can optimize the time resolution of the oscilloscope and the amplitude of the signals, etc.

Analysis of the Speed of Light task according to the Experimental design and investigation rubric:

Rubric item(s)	**Expected performance**
Understand the problem at hand Identify relevant variables Identify appropriate theory/model ***L6_3.1***	Identify the role of the different elements in the setup (source, beam splitter, reflector, focusing lens, detectors, …) and sketch a scheme of the experimental setup. Model the situation it as uniform linear motion of light beams with speed "c" and understand that what you need to measure is

	the difference in time of flight of the two beams due different light paths. Correctly identify the light path of each beam.
[Make a testable hypothesis] Design reliable experiment ***L6_3.1***	Understand that you need long enough distances (light paths of at least a few meters) to get a measurable delay. Decide how you will get the result (Single measurement or repeated measurement? Single distance or multiple distance? How will you calculate the best value, e.g., linear fit?).
Identify appropriate equipment ***L6_3.2***	Understand what is visualized on the oscilloscope screen and how you measure the time difference between the two signals. Optimize the visualization. Adjust the source and detectors parameters (e.g., decide which LED to use; set the pulse width and period of the source if not already given; decide whether to amplify each of the two detected signals).
Construct/assemble experiment Optimization ***L6_3.2***	Correctly make all the electronic connections (power the source, the detectors, and the oscilloscope; connect oscilloscope channels to the detectors; connect oscilloscope to the PC, launch the software and obtain a signal). Evaluate if what is seen by the detectors is the actual delayed signal or a signal directly from the source. Adopt appropriate solutions to minimize signals directly from the source (e.g., using black paper to screen the source to the second detector) or from other sources in the room. Adjust the position of the lenses to maximize intensity.
Conduct experiment(s) Collect data ***L6_3.3***	Decide the number of triggers over which to average the measurement. Evaluate the shape of the signal and decide where to put the cursors to minimize uncertainties in the measurement due to the positioning of the cursors. Decide how many distances you need and how many times to repeat the measurement at each distance. Take all the measurements.
Process data ***L6_3.4***	Calculate mean, std dev, and error on mean for measurements at each distance; plot the data appropriately and make the linear fit to extract the best value for the speed of light.
Present data Comparisons with model, theory, and/or existing data ***L6_3.4***	Correctly display uncertainties associated with data points; compare the result with the known value for the speed of light and discuss the goodness of the result. Discuss possible sources of systematic errors. If time permits, compare for example with different choices of the LED.
Documentation ***L6_3.5***	Record all the data appropriately; record the used settings and any other relevant choices or issues in a logbook; produce a report.
Safety ***L6_3.6***	Operate equipment safely.

Communication prompt related to Dimension 3 tasks:
Present your experimental results to science experts visiting the lab.

Dimension 4
In the next pages, we report the “Problem solving rubric” developed for assessing Dimension 4. Also in this case, the matching with the revised CALOHEE ARF subdimensions is indicated.

PHYSICS

PROBLEM SOLVING RUBRIC				
INDICATOR	**0**	**1**	**2**	**3**
L6_4.1 **Problem framing**	C6_4.1 Identify the physics involved in a problem or situation, identify common structures in different problems.			
4.1.1 **Identification of relevant variables and constraints**	The description is absent or completely erroneous.	The description is incomplete or contains some errors.	The description is useful, but it contains some omissions, assumptions and constraints are not set explicitly.	The description is useful, appropriate, and complete. Evidence of listing assumptions/constraints.
L6_4.2 **Modeling**	C6_4.2 Model the problem mathematically, using different representations (graphs, diagrams, equations) in a consistent way.			
4.2.1 **Representation of the system**	There is not attempt of representing the system o the representation is completely erroneous.	The representation lacks consistency.	Some of the important aspects of the system are missing.	The important aspects of the system are present and represented in a consistent way (field lines, graphs, …).
4.2.2 **Identification of relationships between variables**	All the relationships between the variables are missing or are mistaken.	The relevant relationships between the variables are not identified or are incorrect.	Some of the relationships between the variables are not identified or properly used	All the relevant relationships between the variables are identified (in the example: rate out depends on the height of water).
4.2.3 **Mathematization**	The mathematical concepts necessary for the model is missing, or it is incorrect.	The physical-mathematical model describing the phenomena is not appropriate.	The physical-mathematical model describing the phenomena contains minor omissions or it is flawed.	The physical-mathematical model describing the physical phenomena is appropriate and complete.
L6_4.3 **Execution**	C6_4.3 Choose an appropriate and effective mathematical procedure to solve a problem and execute it in a correct and complete way, justifying any assumptions and simplifications.			
4.3.1 **Execution of mathematical procedures**	Mathematical procedures are missing or are inappropriate.	Some of the mathematical procedures are used incorrectly and/or contains errors.	Mathematical tools and procedures are not properly used with minor omissions and errors.	The mathematical tools and procedures are appropriate and complete.
L6_4.4 **Evaluation**	C6_4.4 Critically evaluate the solution of a problem, discussing its physical meaning and including limiting cases when applicable.			
4.4.1 **Evaluation of the answer**	There is no attempt of evaluating the validity of the solution.	There is an attempt of verifying the validity of the solution, but it is superficial or incorrect.	There is an attempt of verifying the validity of the solution, but it is incomplete or inconclusive.	There is evidence of verification of the validity of the solution (order of magnitude, units, discussion of limiting cases, …).

L6_4.5 **Revision of the process and creative thinking**	C6_4.5 Recognize ways to improve/extend the validity of a solution, devise alternative ways to address a problem and critically evaluate them to choose the best one.			
4.5.1 **Revision of the process and creative thinking**	The entire solution is unclear, unfocused and/or inconsistent.	The problem solution is confused and/or some inconsistencies are present.	The problem solution is logically connected but there is some deficiency or flaw.	The entire problem solution is clear, well focused, logically connected, and alternative ways to look at the same problem are pointed out.

Authentic performance tasks for Dimension 4

In the following we report two examples of problems suitable for the assessment of Dimension 4. Further examples of problem-solving tasks can be found for instance in Burkholder et al., 2020.

The Water Tank problem

To highlight the difference between problems evaluating the "skills" column and the "autonomy and responsibility" column, we developed two versions of this problem, one with a lower level of authenticity (idealized situation) and one with a higher level of authenticity (real-world situation).

Low authenticity version: *A water tank is open at the top and has a hole in the side at the bottom of the tank through which water can flow. Water enters the top at a constant rate. Develop an equation for the height of the water level inside the tank as a function of time, taking into account the rates of water flowing in and out.*

Higher authenticity version: *You are part of a team of experts in charge of building a dam. A river feeds water into the basin created by the dam at a rate that can be assumed constant over much of the year. Technical requirements specify that water should not overflow and that the dam should not dry out under normal conditions. Therefore, an outlet should be built into the bottom of the dam. You are asked to develop a model of the water level behind the dam as a function of time, taking into account the rates of water flowing in and out.*

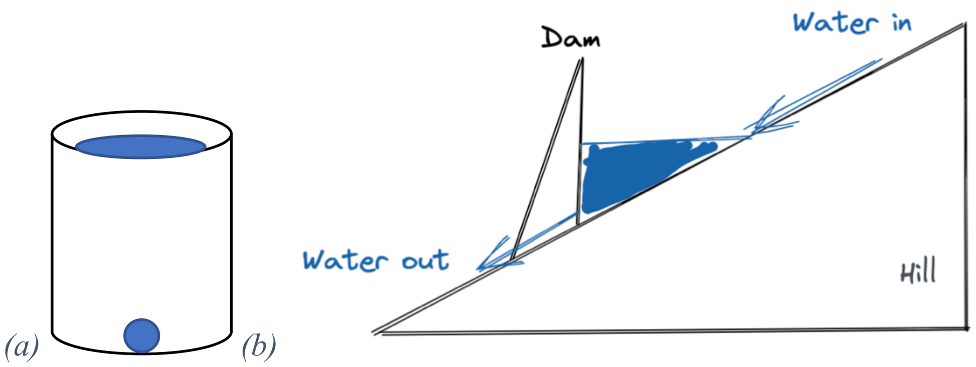


Figure 10 Figures for the Water Tank problem. (a) Low authenticity version. (b) Higher authenticity version.

Analysis of the Water Tank problem (higher authenticity version) according to the Problem solving rubric:

Rubric item(s)	Expected performance
Identification of relevant variables and constraints ***L6_4.1***	Recognition that the important variables are the rates of flow in and out as well as the size of the hole at the bottom.
Representation of the system ***L6_4.2***	Schematic or other sketch of the physical system.
Identification of relationships between variables	Recognition that the rate of change of volume of water in the tank is given directly by the rates of flow in and out; identifying the

L6_4.2	relationship between change in height and change in volume; recognition that the rate out must vary with height of water, otherwise steady state cannot be achieved; the rate out will depend on pressure and therefore depend linearly on the height of the water.
Mathematization ***L6_4.2***	Translation of the relationships above into mathematical form; development of a rate equation by recognizing that rates are given by differentials.
Execution ***L6_4.3***	Solving the differential equation by substitution; development of a logarithmic dependence of height on time.
Evaluation ***L6_4.4***	Does the solution make sense? How does the solution depend on the area of the hole at the bottom? Does the level saturate with time? Steady state is achieved when the rate out is equal to the rate in.
Revision of the process and creative thinking ***L6_4.5***	Examination of the whole process.

The Noise Level problem
For this problem we directly provide the higher authenticity version.

The technicians of your city council want to locate a 6-storey residential building block. The legislation on noise pollution prevention does not allow noise levels above 40 dB inside bedrooms of the dwellings. There are five wind turbines 10 m apart, arranged in line and emitting an acoustic power measured at the base of every wind turbine of 300 Watts each. Power is evenly distributed between the octave bands ranging from 100 Hz to 3000 Hz. The technicians want to know where they can locate the buildings to avoid disturbance, or just in case, what acoustic insulation the material used to build the façade should have. Using appropriate models and with a suitable hypothesis:

1. *Establish the distance from the wind turbines where the dwellings could be located, taking into account air absorption.*
2. *Determine the difference in dB between the noise level inside the bedroom and outside the façade, for the case in which the maximum distance cannot be more than 1 km from the middle of the wind turbines line to the building.*

Data available: Air absorption coefficient tables, octave band table.

Analysis of the Noise Level problem according to the Problem solving rubric:

Rubric item(s)	**Expected performance**
Identification of relevant variables and constraints ***L6_4.1***	Recognition that this is a sound propagation problem. The controlling variables are the acoustic intensity depending on frequency, distance, and energy absorption rate.
Representation of the system ***L6_4.2***	Schematic or other sketch of the physical system. Geometrical representation of the distance location from the line of sound generators.
Identification of relationships between variables ***L6_4.2***	Recognition that the sound intensity depends on distance, and the inverse square law may be applicable if you choose the point source model, and it is feasible for this case. Identifying there is non-coherent interference between sound sources. Identifying that the sound absorption rate is proportional to the sound wave intensity.

	Identifying the relationship between measured sound power on ground and sound intensity. Recognition that the sound absorption depends on frequency and how it depends. The distance will depend on energy, frequency, and absorption coefficient in a non-linear way.
Mathematization ***L6_4.2***	Translation of the relationships above into mathematical form; development of an intensity equation depending on distance and frequency by recognising that sound intensity depends on the square inverse law on distance and the intensity rate decrease linearly with sound absorption coefficient related to frequency.
Execution ***L6_4.3***	Combining the equations, solve a non-linear equation on distance using numerical tools.
Evaluation ***L6_4.4***	Does the solution make sense? Is it important to consider the height of the wind turbines? Is it relevant to take into account the sound absorption? What would be the effect of taking into account meteorological variables, such as wind, temperature or humidity? Will it be important? What would happen if the sound wave generated interferences in some frequencies? What changes would it be necessary to address the problem? Discussion on the possibility of using line source model of sound propagation instead of point source model.
Revision of the process and creative thinking ***L6_4.5***	Examination of the whole process and the use of hypothesis to simplify the problem in the different steps.

Communication prompt related to Dimension 4 tasks:
Present your findings to the general public.

Dimension 7

In the next pages we report the “Communication in physics rubric” that we have developed for assessing communication tasks.

COMMUNICATION IN PHYSICS RUBRIC				
INDICATOR	**0**	**1**	**2**	**3**
L6_7.1 **Information sources**	C6_7.1 Retrieve and use appropriate and reliable sources of information about a scientific topic. Support an argument using reliable sources, also concerning societal issues.			
7.1.1 **Reliability**	The used sources are not reliable.	Few of the used sources are reliable, the number of sources is limited.	Most of the used sources are reliable, but the number of sources is too limited.	The information is retrieved from reliable sources (e.g., publications from recognized journals and/or institutes, master/PhD theses, recognized experts in the field). If possible, more than one source is used to make or support an argument.
7.1.2 **Correct use of the information sources**	The sources do not support the argument.	The sources support the argument, but not in a convincing way.	Most of the sources support the argument.	All the sources are relevant and support the argument in a convincing way.
L6_7.2 **Means of communication**	C6_7.2 Identify and choose the appropriate written or oral style according to the context, audience, and communicative goals (laboratory report, dissertation, scientific article, etc.).			
7.2.1 **Selection of adequate means of communication**	The chosen method of communication is not adequate for the specific audience: the language is too difficult or too simple; the images are not fit, too many or not enough.	The chosen method of communication is not completely adequate for the specific audience. Also, the language is not optimal for the public.	The chosen method of communication is adequate for the specific audience. However, the language is not optimal for the public.	The chosen writing and oral style is tailored perfectly for the specific audience, context and communicative goals.
7.2.2 **Criteria**	The criteria of the presentation are not met.	Some criteria of the presentation are not met.	Almost all criteria of the presentation are not met.	The presentation (oral presentation, report, poster, …) meets the criteria given by the task (e.g., length of the

				presentation and/or the time frame and/or the used media…).
L6_7.3 **Language including scientific English**	C6_7.3 Communicate physics topics using appropriate terms both in own language and in English. Autonomously study physics topics in English, e.g., understanding and reporting about a scientific paper.			
7.3.1 **Language (including scientific English)**	The report is not well structured and concise, and not clear at all. Too many spelling/grammar mistakes are made.	The report is fairly structured and concise but could be improved. There are spelling/grammar mistakes.	The report is structured and concise. There are a few grammar/spelling mistakes.	The language (including scientific English) is used in a concise and structured way. The report is well structured and very clear. There are hardly any language mistakes.

5. Future assessments: where can we go from here

The assessments produced in this work are rather traditional in their format; the reflections of the SAG focused mainly on capitalizing on and integrating best practices from the literature and/or the experience of the SAG members' universities, and on aligning qualifications, assessment, and tasks. In particular, the proposed assessments are thought for in-presence testing and involve manual evaluation from the instructor. Automatically evaluating constructed response assessments or performance-based assessment is, in fact, a challenge recognized by the literature, to which no simple and establish solutions exists yet.

There are, however, some pioneering studies that can provide hints of how emerging technologies could help, in the future, making complex assessments suitable for automated testing. The synthesis reported here is based primarily on the works by Zhai et al. (2020) for a review of studies in science education. Specific examples for physics can be found in Nakamura et al. (2016) and Lee et al. (2019).
Most of the examples reported in the literature rely on machine learning (ML), an emergent computerised technology that relies on algorithms built by 'learning' from training data. Most of the existing studies involve text recognition, classification, and scoring with an emphasis on constructing scientific explanations. As reported by Zhai et al. (2020), many of the present studies are focused on assessing the validity of such systems, while more work is needed to improve the pedagogical and technological aspects. In particular, still a lot of human work is needed to train computers to accurately assess students' work. However, ML shows a potential for assessing complex competences which may also constitute a pedagogical potential for formative assessment. In Figure 11 we report some examples from science education as reported by Zhai et al. (2020).

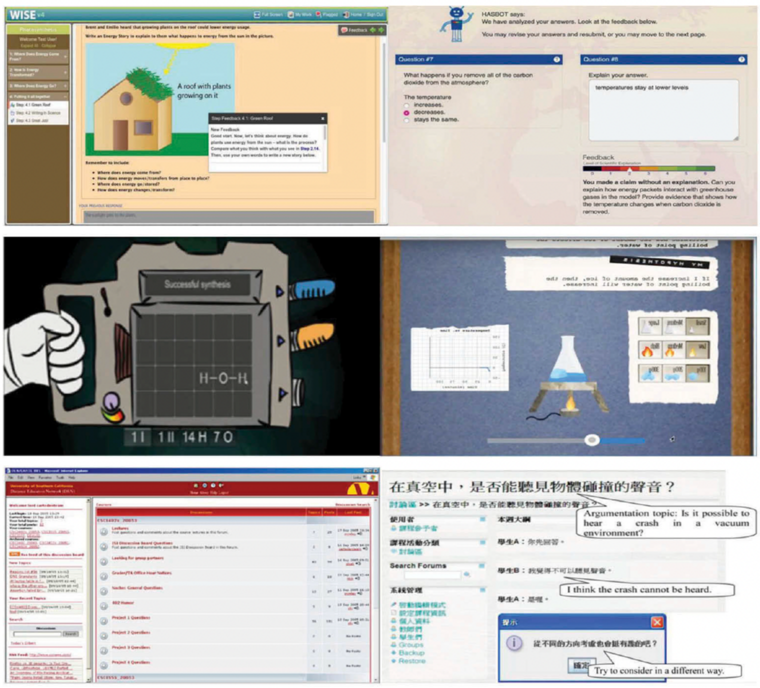


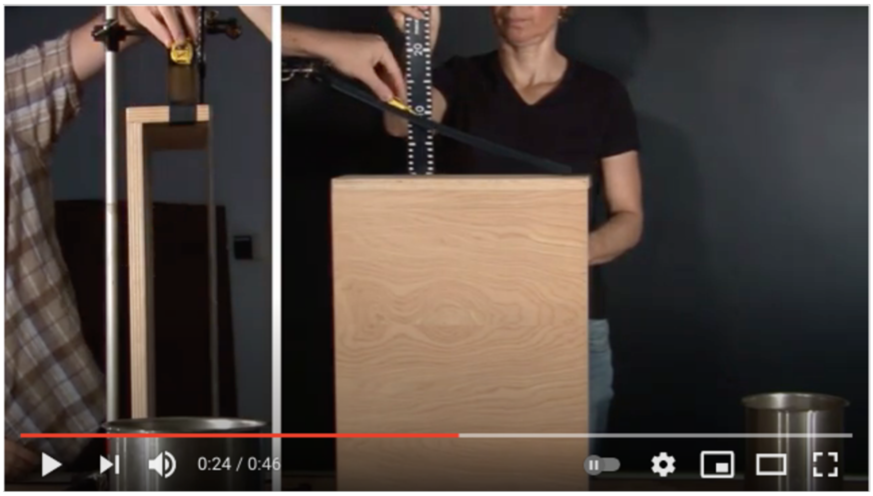
Figure 11 Examples of machine learning-based assessments as reported in Zhai et al. (2020).

6. Actual Testing – Practical Implications

We conclude with some further suggestions about the possible practical use of the proposed assessments.

The first practical application we recommend is that universities use the proposed assessment to evaluate their own courses. The tasks could be proposed to (bachelor) students at the end of the program, serving both as a self-assessment task for students, and as a reference for universities to see if their programmes actually fulfil internationally agree standards for physics competence. Proposing the assessment locally rather than as a single large-scale assessment could make it more practically applicable and still constitute a quality benchmark for the programme.

Another reflection regards the possibility of moving the assessment online. Even though the evaluation remains manual, this option may allow for distance assessment and be more practical to implement. In fact, alongside it many challenges and difficulties, the 2020 pandemic has also fostered the reflection about how to bring experiences that were traditionally proposed only in presence to an online setting. An eminent example is laboratory work. Different solutions have been proposed; here we report an example from the ISLE approach, which has inspired our work on Dimension 3. The proposed solution involves specifically constructed video experiments that are accompanied by lab worksheets adapted from the “classical” ISLE worksheets (Figure 12). The experiments are not self-explanatory but require students to engage in a variety of scientific abilities that can then be evaluated with the usual rubrics. The videos are constructed so that students can extract data from the video itself.



2. Observational experiment: car on a ramp

Goals: a) to apply ideas of work and energy to analyze real life processes;

b) to find and explain patterns in a set of experiments.

Equipment: none

Rubrics for self-assessment: Ability to conduct an observational experiment, B5; B8, B9.

Watch the video of the cart going down the ramp and landing in a can [https://youtu.be/TXJrlOzyYls].

a. Describe what happens in the video when the steepness of the track is changed. What physical quantities do not change? How can you explain the results of the experiment?

Figure 12. Example of an ISLE video experiment.

Although this type of assessment does not cover all of the subdimensions of the CALOHEE Framework and CALOHE2 Rubrics (for example, considerations about the construction of the experiment and safety issues are not applicable), it can be used to evaluate at least a subset of them.

The third and last possibility we suggest is inspired by the PLIC assessment described above (Walsh et al., 2019). In that case, a scenario approach was used, and the automatized answers were constructed moving from typical students' responses observed in the classroom. A similar approach could be used by universities using the CALOHE2 assessments. After the CALOHE2 tasks have been proposed to some cohorts of students, their typical answers could be coded and transformed – at least partially – in closed answers that can be delivered using a computer. This solution still requires a significant amount of work but is technologically ready and practicable.

References

Docktor, J.L., Dornfeld, J., Frodermann, E., Heller, K., Hsu, L., Jackson, K.A., Mason, A., Ryan, Q.X., and Yang, J. (2016). Assessing student written problem solutions: A problem-solving rubric with application to introductory physics. *Phys. Rev. Phys. Educ. Res.* 12(1), 010130. Doi: 10.1103/PhysRevPhysEducRes.12.010130.

Etkina, E., Van Heuvelen, A., White-Brahmia, S., Brookes, D.T., Gentile, M., Murthy, S., Rosengrant, D., and Warren, A. (2006). Scientific abilities and their assessment. *Phys. Rev. S. T. – Phys. Educ. Res.* 2, 020103. Doi: 10.1103/PhysRevSTPER.2.020103.

Lee, H.S., Pallant, A., Pryputniewicz, S., Lord, T., Mulholland, M., and Liu, O.L. (2019). Automated text scoring and real-time adjustable feedback: Supporting revision of scientific arguments involving uncertainty. *Sci. Educ.* 103(5), 590–622. Doi: 10.1002/sce.21504.

Laverty, J.T., Underwood, S.M., Matz, R.L., Posey, L.A., Carmel, J.H., Caballero, M.D., et al. (2016). Characterizing College Science Assessments: The Three-Dimensional Learning Assessment Protocol. *PLoS ONE* 11(9), e0162333. Doi: 10.1371/journal.pone.0162333.

Nakamura, C.M., Murphy, S.K., Christel, M.G., Stevens, S.M., and Zollman, D.A. (2016). Automated analysis of short responses in an interactive synthetic tutoring system for introductory physics. *Phys. Rev. Phys. Educ. Res.* 12, 010122. Doi: 10.1103/PhysRevPhysEducRes.12.010122.

Pols, F. (2020). A Physics Lab Course in Times of COVID-19. *EJRSME 24*(2), 172-178.

Walsh, C., Quinn, K.N., Wieman, C., and Holmes, N.G. (2019). Quantifying critical thinking: Development and validation of the physics lab inventory of critical thinking. *Phys. Rev. Phys. Educ. Res.* 15, 010135. Doi: 10.1103/PhysRevPhysEducRes.15.010135.

Wattanakasiwich, P., Taleab, P., Sharma, M., and Johnston, I. (2013). Development and Implementation of a Conceptual Survey in Thermodynamics. *IJISME* 21(1), 29-53.

Zhai, X., Yin, Y., Pellegrino, J.W., Haudek, K.C., and Shi, L. (2020). Applying machine learning in science assessment: a systematic review. *Stud. Sci. Educ.* 56(1), 111-151. Doi: 10.1080/03057267.2020.1735757.