education sciences

MDPI

Review

Take-Home Exams in Higher Education: A Systematic Review

Lars Bengtsson

Physics Department, University of Gothenburg, SE-405 30 Gothenburg, Sweden; lars.bengtsson@physics.gu.se

Received: 19 September 2019; Accepted: 1 November 2019; Published: 6 November 2019

Abstract: This work describes a systematic review of the research on take-home exams in tertiary education. It was found that there is some disagreement in the community about the virtues of take-home exams but also a lot of agreement. It is concluded that take-home exams may be the preferred choice of assessment method on the higher taxonomy levels because they promote higher-order thinking skills and allow time for reflection. They are also more consonant with constructive alignment theories and turn the assessment into a learning activity. Due to the obvious risk of unethical student behavior, take-home exams are not recommended on the lowest taxonomy level. It is concluded that there is still a lot of research missing concerning take-home exams in higher education and some of this research may be urgent due to the emergence of massive online open courses (MOOCs) and online universities where non-proctored exams prevail.

Keywords: take-home exam; in-class exam; higher-order cognitive skills; unethical student behavior

1. Introduction

Assessment is a necessary part of academic studies on all levels. With few exceptions, an in-class, closed-book, invigilated pen-and-paper exam is the traditional assessment method [1]. There are certainly other assessment methods in use, but the main assessment method at prominent universities is still a proctored, in-class exam (ICE). ICEs are typically characterized by hard time limits (2–6 h) and the stress this imposes on the students. The main reason for advocating ICEs seems to be that it minimizes the risks of the exam being compromised by unethical student behavior [1,2], but it has been criticized for several reasons: it deludes students to superficial learning [1], it does not promote students' 'generic skills' [3], it imposes an unnatural pressure on the students that has an adverse impact on their performance [4], it is not consonant with the prevailing theory of 'constructive alignment' in higher education [1,5,6] and it is not suitable for assessing students' performance on the higher levels of Bloom's taxonomy scale [7–9]. Bloom's taxonomy scale [9] is a hierarchical description of students' learning (revised by Anderson et al. in 2001 [7]). This taxonomy comprises all learning domains (cognitive, affective and sensory), but in this context (as in most higher education contexts) we are only considering the cognitive domain. At the lowest level, students' learning is characterized by root learning ('remember'). The succeeding levels are 'understanding', 'applying', 'analyzing', 'evaluating' and 'creating'; students move from root learners to true scholars where they create knew knowledge. The idea of Bloom's taxonomy is that it describes what phases learning undergoes and as teachers, it is paramount to understand on what level the present students are since this has direct consequences for the curriculum (objectives and activities), but, most of all, it has direct consequences for the design of the exam. An in-class, multiple-choice test (MCQ) may be justified on the lowest levels, but may not be appropriate on the highest levels; the higher levels require that students can define problems, predict, hypothesize, experiment, analyze, conclude and are capable of reflective thinking [10] and they also indicate an "intrinsic creativity or an ability to express ideas in their own

words" [11] (p. 57). These skills can only (or preferably) be tested with ill-defined/ill-structured or open-ended questions [12] which require an abundance of time to answer that is not facilitated by ICEs. These are forcible reasons for researching/reviewing take-home exams (THEs). Further, there appears to be no academic literature to support the widespread use of ICEs as the predominant assessment method [1].

There are also other reasons to dispute the use of ICEs. During the last couple of decades, two dramatic events have fundamentally changed the basis on which universities conduct tertiary education; the massification of higher education [13] and the emergence of the Internet. The consequence of the former is that classes grow (both in size and heterogeneity) and the latter has created a wealth of massive online open courses (MOOCs) and e-universities and almost any information is conveniently available from a computer keyboard. Both these events have forced universities to make comprehensive readjustments. The new cohort of students have different study habits and a much wider spread in academic ability and skills when they enter the tertiary training. The MOOCs and e-universities can usually not facilitate ICEs for practical reasons because their students are geographically scattered across the globe [2]. Aggarwal stated already in 2003 that "technological advancements and student demands" have necessitated a shift from a "brick and mortar synchronous environment" to a "click and learn asynchronous environment" [14] (p. 1). This implies a shift from learning on the universities' terms to learning on the students' terms, or as Hall phrased it: "take-home exams fit the new millennium student's lifestyle" [15] (p. 56). Universities also need to meet the stakeholders' (i.e., the employers') expectations and demands; what knowledge and skills are they really looking for in their next recruitment? Someone who can solve a problem under pressure in a couple of hours or someone who can retrieve, apply and synthesize information from the Internet?

There are also some issues with ICEs when it comes to testing of learning objectives. A tertiary course at a university is scaffolded by a formal syllabus that constitutes the framework of the course. The heart of the syllabus is the section that lists the learning objectives. The most fundamental idea about conducting an exam is to confirm that students have met the learning objectives. A 'pass' grade indicates that the student has reached all the course's objectives (?). The problem with an ICE, at least in 'hard' disciplines like STEM (Science, Technology, Engineering and Mathematics), is that due to the imposed time constraint, items that test all objectives cannot be included in an in-class exam (ICE). Even if they are, a student may only need 50% of the maximum score to get a 'pass' grade. Hence, a student who is awarded only a pass grade has most likely not reached all the objectives in the syllabus. This could be effectively neutralized by introducing THEs (take-home exams). Due to the elongated time-span and the unlimited access to information, there are no longer any reasons not to include test items that cover all syllabus' learning objectives and to pass the exam, students must prove that they have met all objectives. This would be a reasonable consequence of introducing THEs in STEM disciplines—it would benefit students' learning and would be a ponderous argument against incredulous stakeholders.

This work emerged from a need to find alternative assessment methods consonant with the 2020 students' study habits, e-universities' conditions, future stakeholders' demands and to see if a general introduction of THEs in STEM disciplines could be justified. For these reasons, the research concerning the traditional take-home exam (THE) has been systematically reviewed. In this context, a THE is an exam that the students can do at any location of their choice, it is non-proctored and the time limit is extended to days (rather than hours as is the typical time limit for an ICE). As opposed to 'home assignments', THEs are always 'high-stake', i.e., they have a decisive impact on the students' grade. THEs have been used for decades but mostly for certain disciplines; they are well established in 'soft' disciplines (like psychology) but less common in 'hard' disciplines (like medicine and engineering). One of the aims of this work was to review the hitherto research conducted on THEs and to understand why they are prevalent in some disciplines while unwonted in others. THEs have some apparent advantages and disadvantages. The extended time limit implies less stress on the students, more complex and open-ended questions can be used which would increase the test

liability. On the disadvantage side, there is of course the apparent risk of cheating when the exam is not proctored. One objective of this review was to map the community's consensus concerning non-proctored take-home exams, contrast the advantages with the disadvantages and thereby provide a basis for deciding if, when and how to implement THEs in higher education. This should be of uttermost interest to tertiary educators in general and to those who conduct their courses online in particular (where in-class exams may be inconvenient or impractical). Another objective was to find scientific arguments that could justify a spread of THEs (take-home exams) also into STEM disciplines.

To that end, the following questions were targeted:

Q1: What advantages and disadvantages of THEs are contended by the community?
Q2: What are the risks of THEs and can they be mitigated enough to warrant a wide-spread use?
Q3: Are THEs only appropriate for certain levels on Bloom's taxonomy scale?
Q4: THEs are non-proctored, students have access to the Internet and the time-span is (typically) extended. How does that affect the question items on the THE?
Q5: How do THEs affect the students' study habits during the weeks preceding the exam and how does that affect their long-term retention of knowledge?
Q6: Do THEs promote students' higher-order cognitive skills (HOCS)?

By HOCS, we refer to the ability to find, validate, select, integrate, synthesize, communicate, comprehend and present information, to function in teams, critical thinking, ethical responsibility, sustainability and social commitments and the ability to provide a holistic perspective [3,16]. 'Retention of knowledge' is defined as the knowledge that students retain (some) weeks after the initial testing [17].

This work is closely related to a previous review conducted by Durning et al. [18] where they reviewed the research to date on open-book (OBE) and closed-book (CBE) exams. The work conducted here differs from Durning et al.'s work in that their OBEs were still proctored (unethical behavior was not considered) and their focus was mainly medical students. This work focused specifically on non-proctored THEs and its implications/issues (and included all disciplines), but THEs are closely related to open-book exams; THEs are just an extension of OBEs [15] and some results from OBE research will be pertinent to this review. For example, despite the fact that several works unequivocally prove the correlation between high performance on ICEs and better practice outcomes [19], Durning et al. were not able to conclude that ICEs should be the preferred examination method in health care professions [18].

This work was, for the most part, a qualitative review, focusing on the occurrence of the themes and conclusions rather than on quantitative results (analysis).

2. Methods

This systematic review was conducted according to Gough's nine-phase process [20] outlined by Bearman et al. [21]. This process stipulates that database searches are preceded by elaborate inclusion/exclusion criteria as well as articulated search and screening strategies. Potential works were primarily identified by searching five databases (Education Database, ERC, ERIC, Scopus and Web of Science) with characteristic search phrases and pursuing cited references in these works (both forward and backwards). These searches were finally complemented with a search on Google Scholar.

2.1. Keywords

The primary keyword was 'take-home exam' restricted to 'higher education'. Databases' thesauruses suggested that 'test' and 'assessment' are valid synonyms to 'exam', and 'tertiary education' is a valid synonym to 'higher education'. Hence, in the five primary databases, the following search condition was used:

'take-home exam*' OR 'take-home test*' OR 'take-home assessm*'—title/abstract/keywords
AND

'higher education' OR 'tertiary education'—all fields

In Google Scholar, only 'take-home exam*' AND 'higher education' was used in all fields, but patents and citations were excluded. Table 1 illustrates the number of hits produced by each database.

Table 1. Number of hits in each database.

Database	Hits
Education database	474
Education research complete	22
ERIC	33
Scopus	17
Web of Science	3
Google Scholar	1010

2.2. Screening Algorithm

The screening and inclusion algorithm is illustrated in Figure 1.

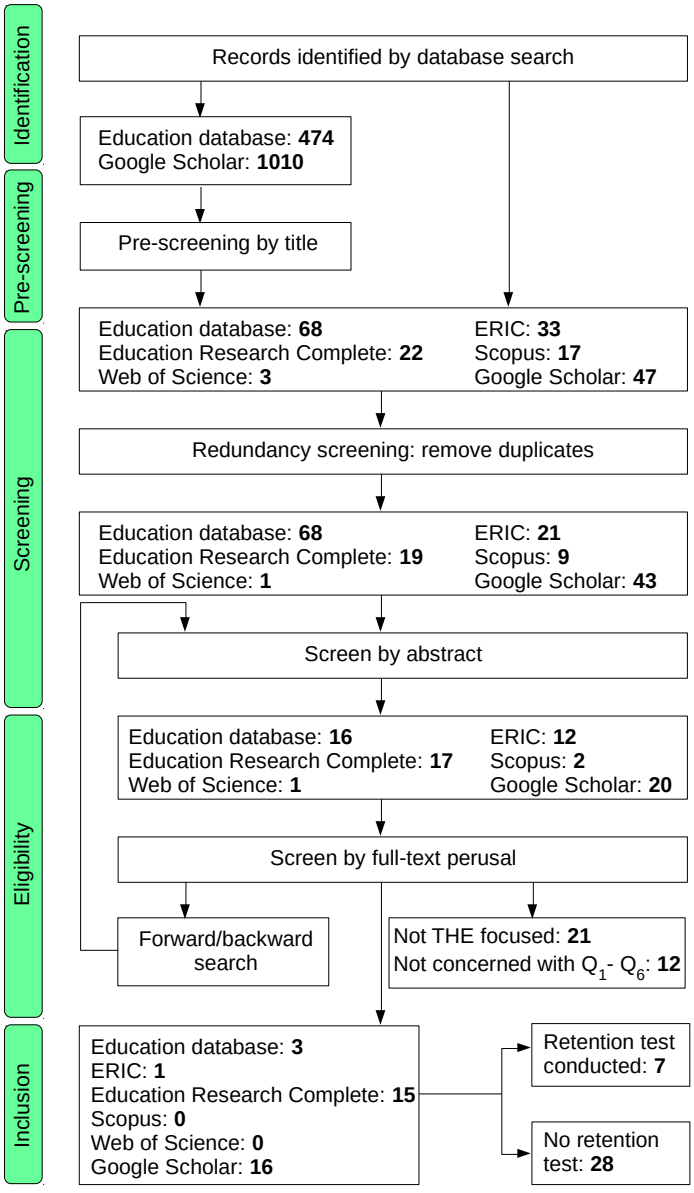

Figure 1. Screening process algorithm.

No screening for subjects or students' level was applied; both undergraduates and postgraduates were included (colleges, universities, vocational schools etc.). There was also no screening for type of tertiary education and no geographic preferences were applied.

The huge number of hits in Education Database and Google Scholar called for a pre-screening process where only titles were used to determine the relevance to this review. This was followed by a redundancy screening where duplicates were removed. Next, data samples were screened by abstracts and the remaining samples were perused full-text. The full-text perusal also included forward and backward 'snowball sampling' for additional items.

2.3. Inclusion/Exclusion Criteria

In the pre-screening process that was applied to the Education Database and Google Scholar samples, the main inclusion criteria were that the title, or the first page excerpt, should include the phrases 'take-home' and 'exam/test/assessment' or otherwise give a clear indication of being related to question items Q_1–Q_6. For example, if the title included keywords like 'HOCS', 'higher-order cognitive skills', 'non-proctored', 'open-book exam', 'retention test' or 'Bloom taxonomy' they were passed forward to the abstract screening stage.

Abstracts from 168 samples were screened by exclusion criteria; if it was obvious that they were not concerned with question items Q_1–Q_6, they were excluded. 68 samples remained at the full-text perusal stage. In order to organize and code the works, a matrix was designed with one column for each question (Q_1–Q_6) and any content in the works that was identified as relevant to any of the questions was copied into the matrix. This facilitated a convenient basis for the subsequent analysis of all works, question by question. After the full-text perusal of all 68 works, the matrix contained 35 works that were considered to significantly contribute to answer question items Q_1–Q_6. Seven works were considered particularly interesting because they supported their conclusions by conducted retention tests.

These 35 works are presented in Table 2, where their contributions to each question item is indicated. The methods used in all these works are summarized in Table 3, where also the validation of each work has been assessed. The main validation criteria were that controlled experiments had been performed on random groups followed by sound statistical analyses (p-values or other quantitative numbers accounted for) (= 'Yes'). 'Personal reflections' and 'Personal comments' indicate that no experiments have been performed; reasoning and conclusions are based on the authors' personal experience only. 'Hypothesis testing' means that at least a null hypothesis has been formulated and properly tested by statistical tools (paired t-tests, ANOVA, etc.) and p-values are properly accounted for ('Yes'). 'Case study' indicate that take-home exams have been implemented but no hypothesis was formulated, and results and conclusions were based on non-quantitative analyses (interviews, mostly). 'Survey' indicate that anonymous questionnaires were used to probe students' opinions/attitudes and 'Review' means that others' works have been summarized. 'Pilot study' refers to an investigation where participation was voluntary and 'Synthesizing from others' means that data from several other experiments were analyzed and new conclusions were reported. In order to be classified as 'Validated' ('Yes'), the experiments must have been properly designed (randomized groups, clear hypothesis/research question), properly analyzed with established quantitative methods.

Table 2. Included works and their contributions to this review (listed by publication year).

Work	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Retention
Freedman, 1968 [22]	√	√		√	√		
Svoboda, 1971 [10]	√	√		√	√	√	
Marsh, 1980 [23]		√	√		√		√
Weber et al., 1983 [24]	√	√	√		√		
Marsh, 1984 [25]		√	√		√		√
Grzelkowski, 1987 [26]				√		√	
Zoller and Ben-Chaim, 1989 [27]	√						
Murray, 1990 [28]	√	√				√	
Fernald and Webster, 1991 [29]	√	√	√	√	√	√	√
Haynie, 1991 [17]	√		√	√	√		√
Andrada and Linden, 1993 [8]	√		√	√	√	√	
Ansell, 1996 [30]						√	
Norcini et al., 1996 [31]	√			√	√		
Hall, 2001 [15]	√	√			√		
Mallory, 2001 [2]		√					
Zoller, 2001 [12]						√	
Haynie, 2003 [32]		√			√		√
Bredon, 2003 [11]		√	√	√			
Tsaparlis and Zoller, 2003 [33]	√		√	√	√	√	
Giordano et al., 2005 [34]	√	√					
Moore and Jensen, 2007 [35]	√				√		
Williams and Wong, 2009 [1]	√	√		√	√	√	
Frein, 2011 [36]		√					
Giammarco, 2011 [6]	√				√		√
Lopez et al., 2011 [3]	√	√	√		√		
Rich, 2011 [4]	√		√		√	√	√
Marcus, 2012 [37]		√					
Tao and Li, 2012 [38]	√	√		√			
Hagström and Scheja, 2014 [39]			√	√		√	
Rich et al., 2014 [40]	√				√		
Sample et al., 2014 [41]						√	
Johnson et al., 2015 [42]	√			√		√	
Downes, 2017 [43]		√					
D'Souza and Siegfeldt, 2017 [44]	√	√					
Lancaster and Clarke, 2017 [45]	√	√					

Table 3. Summary of methods and validation.

Work	Method	Validated
Freedman, 1968 [22]	Personal reflections	No
Svoboda, 1971 [10]	Personal comments based on experience	No
Marsh, 1980 [23]	Hypothesis testing	Yes
Weber et al., 1983 [24]	Hypothesis testing	Yes
Marsh, 1984 [25]	Hypothesis testing	Yes
Grzelkowski, 1987 [26]	Case study	No
Zoller and Ben-Chaim, 1989 [27]	Survey + Case study	Yes
Murray, 1990 [28]	Review	No
Fernald and Webster, 1991 [29]	Case study	No
Haynie, 1991 [17]	Hypothesis testing	Yes
Andrada and Linden, 1993 [8]	Hypothesis testing	Yes
Ansell, 1996 [30]	Collaborative THE (voluntary pairing) + final ICE	No
Norcini et al., 1996 [31]	Hypothesis testing	Yes
Hall, 2001 [15]	Pilot study with optional take-home test	No
Mallory, 2001 [2]	Case study	No
Zoller, 2001 [12]	Case study	No
Haynie, 2003 [32]	Hypothesis testing	Yes
Bredon, 2003 [11]	Multiple choice THE	Yes
Tsaparlis and Zoller, 2003 [33]	Synthesizing from others	Yes
Giordano et al., 2005 [34]	Hypothesis testing	Yes
Moore and Jensen, 2007 [35]	Hypothesis testing	Yes
Williams and Wong, 2009 [1]	Online survey, students reminded by e-mail	No
Frein, 2011 [36]	Hypothesis testing	Yes
Giammarco, 2011 [6]	Hypothesis testing	Yes
Lopez et al., 2011 [3]	Case study	No
Rich, 2011 [4]	Theoretical work	No
Marcus, 2012 [37]	Review	No
Tao and Li, 2012 [38]	Hypothesis testing	Yes
Hagström and Scheja, 2014 [39]	Hypothesis testing	Yes
Rich et al., 2014 [40]	Hypothesis testing	Yes
Sample et al., 2014 [41]	Hypothesis testing	Yes
Johnson et al., 2015 [42]	Case study	Yes
Downes, 2017 [43]	Review	No
D'Souza and Siegfeldt, 2017 [44]	Hypothesis testing	Yes
Lancaster and Clarke, 2017 [45]	Review	No

3. Results

The results are presented as a summary of the conducted research associated with each one of the posed questions.

Q1: What advantages and disadvantages of THEs are contended by the community?

Proposed advantages and disadvantages are listed in Tables 4 and 5 respectively, in order of most frequently cited advantage/disadvantage.

The most cited advantages are the reduction of students' anxiety, the opportunity to test HOCS and conservation of classroom time. The main disadvantage of THEs, purported repeatedly, is the apparent risk of unethical student behavior.

Q2: What are the risks of THEs and can they be mitigated enough to warrant a wide-spread use?

The by far most frequently cited risk of THEs is the apparent risk of unethical student behavior, and a lot of effort has been made to prevent or diminish the risks of cheating. Table 6 summarizes the remedies proposed by the community.

Table 4. Advantages of THEs.

Purported Advantage	Source
Reduce students' anxiety	Fernald and Webster, 1991 [29]; Giordano et al., 2005 [34]; Hall, 2001 [15]; Johnson et al., 2015 [42]; Rich, 2011 [4]; Rich et al., 2014 [40]; Tao and Li, 2012 [38]; Weber, et al., 1983 [24]; Williams and Wong, 2009 [1]; Zoller and Ben-Chaim, 1989 [27]
Can be designed to tests HOCS	Williams and Wong, 2009 [1]; Andrada and Linden, 1993 [8]; Fernald and Webster, 1991 [29]; Freedman, 1968 [22]; Giordano et al., 2005 [34]; Johnson et al., 2015 [42]; Lopez et al., 2011 [3]; Sample et al., 2014 [41]; Svoboda, 1971 [10]; Zoller, 2001 [12]; Zoller and Ben-Chaim, 1989 [27];
Conservation of classroom time	D'Souza and Siegfeldt, 2017 [44]; Fernald and Webster, 1991 [29]; Haynie, 1991 [17]; Svoboda, 1971 [10]; Tao and Li, 2012 [38]; Fernald and Webster, 1991 [29]
Provide a good learning experience	Andrada and Linden, 1993 [8]; Rich et al., 2014 [40]; Zoller and Ben-Chaim, 1989 [27]
Places responsibility for learning on the student	Grzelkowski, 1987 [26]; Hall, 2001 [15]; Lopez, et al., 2011 [3]; Rich, 2011 [4]
Promote learning by testing	Andrada and Linden, 1993 [8]; Freedman, 1968 [22]; Rich, 2011 [4]
Students have time to accomplish tasks that take time	Svoboda, 1971 [10]
Promote a more realistic student study	Svoboda, 1971 [10]; William and Wong, 2009
Less burdensome for teachers	Lopez, et al., 2011 [3]; Svoboda, 1971 [10]
Engage students rather than alienate them	William and Wong, 2009; Zoller and Ben-Chaim, 1989 [27]
Students spend more time on a THE than on an ICE	Freedman, 1968 [22]; Rich, 2011 [4]
Contribute toward a more interactive teacher/student learning environment	Hall, 2001 [15]
Foster the educational process beyond that of memorization	Foley, 1981 [46]
Do not require venue of supervision	Hall, 2001 [15]
Extended involvement of students with course material	Freedman, 1968 [22]
Luck is ruled out	Freedman, 1968 [22]
Offer flexibility regarding location and time	Williams and Wong, 2009 [1]
Bring convenience to both instructor and students	Tao and Li, 2012 [38]
Students learn more and study harder	Rich et al., 2014 [40]
Enforce students to consult other texts apart from the course textbook	Zoller and Ben-Chaim, 1989 [27]
Shift perspective from teaching to learning	Lopez, et al., 2011 [3]
Can assess also the highest Bloom levels	Lopez, et al., 2011 [3]
More sophisticated questions can be asked	Lopez, et al., 2011 [3]
A greater rigor in the answers can be demanded	Lopez, et al., 2011 [3]
Make it possible to assess teamwork	Lopez, et al., 2011 [3]
Questions can be asked about all the content of the syllabus	Lopez, et al., 2011 [3]
Improve retention	Johnson et al., 2015 [42]
Permit testing of content not covered in class/textbook	Johnson et al., 2015 [42]

Table 5. Disadvantages of THEs.

Purported Disadvantage	Source
Easily compromised by unethical student behavior	Bredon, 2003 [11]; Downes, 2017 [43]; D'Souza and Siegfeldt, 2017 [44]; Fernald and Webster, 1991 [29]; Frein, 2012 [36]; Hall, 2001 [15]; Lancaster and Clarke, 2017 [45]; Lopez et al., 2011 [3]; Mallory, 2001 [2]; Marsh, 1980 [23]; Marsh, 1984 [25]; Svoboda, 1971 [10]; Tao and Li, 2012 [38]; William and Wong, 2009 [1]
Students attend fewer lectures	Moore and Jensen, 2007 [35]; Tao and Li, 2012 [38]
Students submit fewer extra-credit assignments	Moore and Jensen, 2007 [35]; Tao and Li, 2012 [38]
Writing and marking is time consuming	Andrada and Linden, 1993 [8]; Hall, 2001 [15]
Students only hunt for answers	Haynie, 1991 [17]
Undermine long-term learning	Moore and Jensen, 2007 [35]
Promote lower levels of academic achievements	Moore and Jensen, 2007 [35]
Items used on THEs are forfeit	Andrada and Linden, 1993 [8]
Generate long answers	Rich et al., 2014 [40]

Table 6. Remedies for unethical student behavior on non-proctored THEs.

Remedy	Source
If questions are designed so that they require a thorough understanding of course material, they will be costly to contract out	Bredon, 2003 [11]
Ask for proof and justifications to all answers	Lopez et al., 2011 [3]; Svoboda, 1971 [10]
Introduce an honor code	Fernald and Webster, 1991 [29]; Frein, 2011 [36]
Grade down for copy without reference	Freedman, 1968 [22]
Submit electronically to permit plagiarism control	Williams and Wong, 2009 [1]
Answers must make direct references to course-specific material	Williams and Wong, 2009 [1]
Make questions "highly contextualized"	Williams and Wong, 2009 [1]
Cohort cheating can be detected by statistical means	D'Souza and Siegfeldt, 2017 [44]; Weber et al., 1983 [24]
Narrow the timeframe to complete the test	Frein, 2011 [36]; Lancaster and Clarke, 2017 [45]
Randomly scramble the order of questions and answers	Frein, 2011 [36]; Tao and Li, 2012 [38]
Use a security browser to prevent printing and saving of the exam	Tao and Li, 2012 [38]
Implement remote invigilation services	Lancaster and Clarke, 2017 [45]; Mallory, 2001 [2]
Assign questions randomly	Murray, 2012 [28]
Ask for hand-written answers	Lopez et al., 2011 [3]
Print exam with watermarks	Lopez et al., 2011 [3]

The cheating issue seems to dominate all conducted risk analyses that have been published, but other concerns have been voiced. There is some concern that students will not read the whole course material but only hunt for answers, but this can easily be mitigated by including questions about all the material [17,32,38] (See also Q5 below).

Q3: Are THEs only appropriate for certain levels on Bloom's taxonomy scale?

No research has been found that specifically addresses this issue, but it has been commented in some works. However, several researchers assert that THEs are more appropriate on the higher taxonomy levels and they also do not recommend them on the lower levels because answers can easily be retrieved from the Internet/textbook or copied from peers [3,4,8,17,24]. Hagström and Scheja [39] took it one step further and proved that introducing a meta-reflection on THEs stimulated a deep approach to learning (students were asked to describe and motivate their strategies and literature used).

Q4: THEs are non-proctored, students have access to the Internet and the time-span is (typically) extended. How does that affect the question items on the THE?

There seems to be an almost unanimous opinion in the community that THE question items should be designed to test the higher taxonomy levels of understanding and/or generic higher-order cognitive skills (HOCS). Questions should be open-ended and require prose or essay answers rather than multiple-choice questions (MCQs) because it is hard to design MCQs that go beyond the memorization level [4,8,10,11,26,31,38,42]. Due to the extended time-span implicated by a THE and the fact that students are "freed from the toil of memorization", "the instructor can get tough for good reasons" [22] (p. 344–345). A THE allows the instructor to include questions covering all the course material [17,32,38]. Questions should force students to higher level thinking, to apply knowledge to novel situations and synthesize material [17]. Bredon [11] exemplifies this by suggesting that graphs could be used with adhering questions that force students to draw interpolating and extrapolating conclusions from graphs. In politics, Svoboda [10] (p. 231) exemplifies this by distinguishing between typical ICE questions, like "What are the three major branches of the federal government?", whereas on a THE, the question should rather be "Should we have governments?" in order to elicit/assess students' higher level thinking skills.

Q5: How do THEs affect the students' study habits during the weeks preceding the exam and how does that affect their long-term retention of knowledge?

This appears to be the most polarized question in the community, both concerning the study habits and the long-term retention. Some researchers assert that students who know that they will be assessed by a THE tend to study less than if they are assessed by an ICE [3,15,23,25,32,35]. Others take the polar, opposite standpoint and argue that they study more if they know they will have a THE [4,10,22,24,33,40]. The main disagreement seems to be whether THEs allure students not to read the entire course material and instead only hunt for answers once the THE is available.

It is interesting to note that the group of researchers that claims that THEs promote long-time retention learning better than ICEs, are (almost) identical to the group who claims that students' study more for THEs. However, only seven works were found where actual retention tests were conducted (as unannounced tests sometime after the THE) to support their claims [4,6,17,23,25,29,32].

Q6: Do THEs promote students' higher order cognitive skills?

The community seems to agree that THEs are an excellent tool for promoting HOCS; THEs both cultivate HOCS [8,10,22,27,41,42] and facilitate a more accurate assessment of HOCS [3]. By including questions from material that is not covered in class, HOCS are cultivated and a shift from algorithmic problem solving to conceptual understanding is fostered [33]. THEs are also recommended for promoting, cultivating and assessing team work [4,10,30,42]. THEs have also been proven to foster an understanding of the learning process [1].

The extended time-span allows students to meta-reflect on their answers which has been proven to have a benign impact on their scoring results [39]. However, designing appropriate question items to assess students' HOCS is a challenging task even for experienced instructors (ibid).

4. Discussion

4.1. Community Consensus

The community seems to agree that there are two major advantages associated with THEs: they reduce students' anxiety and they are an excellent tool when it comes to testing students' higher-order thinking skills. It should be noted though, that the issue of students' stress in examination situations has been debated elsewhere and it has been advocated that some stress is favorable for students' performance [47]. The general opinion that THEs favor HOCS also indicates that they should be appropriate for assessing students' skills on the higher levels of Bloom's taxonomy scale (evaluate, create; even if no research has specifically targeted this issue).

Conservation of classroom time is highly appreciated by the community; more time (and money) can be spent on teaching when proctored exam venues are not required. It is interesting though,

that there is a disagreement in the community about whether THEs take more, or less, teacher time to administer [3,8,10,15,40].

A transfer of the responsibility of learning from the teacher to the student is emphasized. Well, that probably assumes that the student is mature enough to really take that responsibility; it implies a certain maturity, both in academic and generic skills and is consonant with the assumption that THEs are best implemented on the higher taxonomy levels.

4.2. Cheating

The apparent risk of unethical student behavior associated with THEs is the most cited concern. A non-proctored exam conducted in a closed dorm room with an Internet access is the perfect setup for frame-ups and imposture. It could probably be accused of being naïve. Cheating does occur, and not only at the less renown institutions; in 2012, Harvard suffered an immense scandal when nearly half of the students in an introductory course in politics were accused of cheating on a THE [48,49] and similar incidents have been reported from Duke [50,51], West Point [52], Ohio State University [43] and University of Central Florida [37]. In the wake of the Harvard cheating scandal, Stanford University's stakeholders also expressed their concern about the increased use of THEs at Stanford [53]. It also needs to be pointed out that even if illicit collaboration and plagiarism are the two most obvious violations of THE restrictions, 'pens-for-hire' is an increasing business that feeds on non-proctored examinations around the world; all kinds of homework and THEs can be contracted out to online papermills and essay factories [11]. Table 6 lists the community's suggested remedies, but at least some of them could be accused of being naïve; an honor code will probably not deter a lot of offenders. Most of the countermeasures listed in Table 6 suggest that questions should be designed to make the THE hard to contract out; they should require a deep understanding of the material and answers should always be justified by proof, well-founded arguments and direct references to course material (or other sources). Again, this implicates that THEs should be restricted to the higher taxonomy levels.

The cheating issue associated with THEs draws a distinct line between scholar agitators. Some claim that the cheating rate does not increase with THEs [1,11,24], that THEs are no more of a sham than any other form of exams [22] and some scholars even advocate that there might not even be anything wrong with consulting fellow students or other persons, because this is how we would solve any other problem [10]. Others contend that cheating apparently compromises the THE as a credible assessment method; "honesty, as most other variables, is normally distributed" [23] (p. 289). A significant proportion of professors strongly dissuade/dismiss the use of THEs in tertiary education: "Examination without invigilation should not be considered culturally acceptable" [45] (p. 225).

A lot of research has been conducted concerning unethical student behavior. It has been suggested that cheating is more common among students from well-educated families [37]. It has been hypothesized that the reason is that these students are under a lot more pressure from home; cheating is bad, failing is worse [37] (p. 23). It has also been reported that older students are far less prone to cheat than younger students [2].

4.3. THEs and Study Habits

Apart from the cheating issue, there seems to be one other major concern about THEs; how do THEs affect the students' study habits and long-term retention? The community is very polarized in this matter. Some research indicate that it has an adverse impact on their long-term retention [23,25,35], others claim that is has a positive impact on retention [4,29] whereas others report no observed differences in retention between ICEs and THEs [6] (a work on OBEs versus ICEs indicated no difference in retention [54]). Similarly with study habits, some researchers claim that the students study less if they know they will have a THE [15,17,23,24,35] and some claim they will study more [1,4,40].

It has been shown that students spend more time conducting a THE compared to an ICE [3,22] and they therefore concluded that THEs constitute a better learning experience. This could be challenged. Let us assume that students learn more conducting a THE than an ICE. Is it not a little precipitous to

conclude that they therefore also have learned more compared to an ICE-assessed course? Long-term retention learning is benefited by allowing time to assimilate and rehearse. If THEs allure students to defer studying and concentrate their efforts only to conducting the THE, it will inhibit assimilation and long-term retention; the extra time they spend conducting the exam does not necessarily make up for the lack of studying preceding the exam.

4.4. Advocates and Objectors

The most salient observation during this review was that the number of works advocating the use of THEs immensely outnumbers the works discrediting the use of THEs even though ICEs still prevail as the main assessment method in tertiary education. The large number of publications in favor of THEs could be explained by the fact that they represent a change away from the established modus operandi and that always attracts more attention from scholars. Marsh [23,25] and Moore and Jensen [35] are the only works found that explicitly dismiss THEs/OBEs as inferior assessment methods as far as retention learning is concerned. However, these two works were explicitly only concerned with the three lowest levels on Bloom's taxonomy scale.

Marsh [23,25] concluded that the students who took the THE studied less the weeks preceding the exam. Well, if the reason for the differences is that the students who know that they will have a THE study less, then maybe they should be deliberately 'deceived' by not disclosing the exam method in advance. If students study harder for an ICE and learn more during a THE, then a 'surprise THE' at the end of the semester might combine the best of two worlds (but maybe that trick would only work once? Rumors and word of mouth from previous students would probably corrupt that strategy the following year).

4.5. Stakeholders

There is also the issue about the stakeholders' continuous trust in our assessment system. Is perhaps the students' unreserved support for THEs [1,27,29] based on a prevailing sentiment that it would simply make their life easier, as you can always pass a THE? This raises an interesting question: what are the students' main concern? If students had to make a choice between an assessment system that renders them a high grade but low retention and another system that renders them a lower grade but higher retention, which one would they chose? As university teachers, we would like to think that they would all chose the high retention system, but that is most likely naïve. A high degree of retention is good, but a low grade could ruin your career. High grades and degrees/diplomas open career opportunities. Students know this, high grades beat high retention every day of the week and that could explain why THEs are favored by students; they think it increases their chances for high grades.

Which brings us to the secondary stakeholders—the future employers. THEs are not (yet) a generally accepted assessment method in 'hard' disciplines like STEM. Due to changes in students' attitudes and study habits and the emergence of geographically scattered students in virtual classrooms (facilitated by Internet infrastructure expansions), the need to introduce THEs on a broad scale in these disciplines may be inevitable. Universities may not oppose; if they can reach more students, they can make more money. The major question is whether non-proctored exams will reduce the value of the degrees we award in the eyes of the customers' (the employers').

Also, the long-term consequences of students not engaging in deep learning but only reverting to chasing high grades will most likely have an adverse impact on the development of their higher-order cognitive skills and obstruct future advancement on Bloom's taxonomy scale.

5. Conclusions

This review indicates that THEs may not be appropriate on the lowest levels of Bloom's taxonomy scale. The opportunity of cheating is simply too tantalizing because the test items on the 'knowledge' level are usually fact oriented and too easily available on the Internet. There is a dispute in the community about whether ICEs or THEs best facilitate deep learning, but this review indicates that

for the higher taxonomy levels, THEs are preferred by the community because higher-order thinking, and reflections, require more time (and less stress imposed on the students). The community seems also to agree that on the higher taxonomy levels, cheating in THEs is a minor problem that can be mitigated/prohibited/detected (or is non-existing).

If THEs are used on the lower taxonomy levels, a combination of ICEs and THEs might be worth considering. This was first suggested by Ebel in 1972 [55] as a means to base the assessment on both the "quickness of intellect" (captured by the ICE) and the "persistence of effort" (captured by the THE). This idea was the basis of an assessment method developed by John Bailey at Clark State University [28]. Bailey used THEs as "a second chance to learn". An ICE is complemented with a THE; when the ICE was handed in, it was exchanged for a second THE. The total score was calculated according to an elaborate formula. Assume that the total score of the ICE is T points and that a student scores IC points on the ICE and TH percent on the THE. The total score is then [28]

$$Total\ Score\ =\ IC + \frac{1}{2} \times (T - IC) \times \frac{TH}{100} \quad (1)$$

If T = 60 points and a student scores 36 points on the ICE and 75% on the THE, the total score is 45 points (36 + 0.5 × 24 × 0.75). According to Bailey [28], this offers students a second chance without "giving away" grades. NB, the score on the ICE was not disclosed until the THE was returned. This "forces" the students to work hard also on the THE and really turn it in on time. Foley [46] suggested a similar approach where the ICE is first returned with only the total score marked. The students were invited (but not required) to take home the ICE and redo it using any aid. If the student can identify wrong answers and provide convincing rationale for it, one can gain credit.

This may be a very good way to force the students to study hard the weeks preceding the exam (ICE) and, also turn the exam into a learning activity (THE). The dispute about whether the ICE or the THE best promote retention learning would be defused because students do both. This method would probably not affect the trust of our stakeholders (but there is a great chance that it would require more teachers' hours for grading two exams) and also smoothly introduce the students to the change in assessment methods awaiting them later as they climb Bloom's taxonomy ladder.

Recommendations

If THEs are used, consider using them on the higher levels of Bloom's taxonomy only and, if possible, do not disclose the exam method in advance (Weber et al. [24] announced the exam method two days in advance and recommended not to disclose it "until last possible moment"). If there are any concerns about the validity of THEs, Bailey's assessment design with a combination of an ICE and a THE is recommended [28].

This review has identified some missing research concerning THEs. Some of this research might be urgent due to the emergence of MOOCs and online e-universities where non-proctored exams prevail.

First, more tests should be conducted where ICE/THE comparisons are supported with delayed retention tests. We suggest that experiments should also investigate if students should know whether they will have an ICE or a THE in advance. If so, when should that be disclosed? Exactly when should the retention test be performed?

Second, the gain in students' HOCS has been widely purported as one of the major advantages of THEs, but hard evidence is scarce. An experiment that contrasts the gains in HOCs between THEs and ICEs is desirable to provide scientific evidence to support the endorsement of THEs.

Third, what is exactly the prevailing attitude towards THEs among the faculty staff? A lot of works have been published that indicate a strong endorsement among students [1,27,29] but the (assumed) resistance among faculty professors has mostly been insinuated [1] and it would enrich the discourse if we could put a 'number on the resistance'. Most of all though, the published works advocating THEs so outnumber the works dissuading from THEs which contradicts the general opinion that most professors are against it. It could be inferred that the 'against' group is underrepresented in

scientific literature and interviewing a (large) number of (random) professors (at different faculties) would facilitate a deeper understanding of the problems attributed to THEs.

There is also no work that has considered the 'secondary' stakeholders' (employers) opinion on this matter; their opinions about THEs are important inputs to the discourse.

Marsh concluded already in 1984 [25] (p. 111) that "there is a paucity of specific literature comparing take-home exams and in-class exams" and [24] (p. 474) came to the same conclusion: "virtually no research is available on take-home examinations". Haynie draw the same conclusion in 1991 [17]. This review reveals that this is still true 28 years later. However, in the light of the emergence of the Internet and its implications for higher education (MOOCs and online e-universities), the need for more research is even more urgent now.

The virtues of THEs have been widely extolled as promoting HOCS and simultaneously provide means to assess students and constitute an additional learning activity [3,17,29]. Everybody recognizes the risk of unethical student behavior but there is a salient difference in attitudes towards the occurrence of cheating and the need for countermeasures. Some claim that the cheating cohort is relatively small and should be more or less ignored: "the main priority should be to focus on the higher quality learning outcomes of the majority, rather than set up an entire system to stop a small minority" [1] (p. 234). This may be underestimating the problem; Lancaster and Clarke collected 30,000 contract cheating requests from students in a recent survey [45]. Careless use of THEs may compromise and devaluate higher education degrees. On the other hand, proper use has the potential to both promote and assess the highest taxonomy levels and foster higher-order cognitive skills.

Finally, it should be pointed out that the 35 works reviewed in this work (Table 2) stem predominantly from Anglo-Saxon contexts and this may introduce a bias; conclusions may very well be applicable to that context only.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Williams, B.J.; Wong, A. The efficacy of final examination: A comparative study of closed-book, invigilated exams and open-book, open-web exams. *Br. J. Educ. Technol.* **2009**, *40*, 227–236. [CrossRef]
2. Mallroy, J. Adequate Testing and Evaluation of On-Line Learners. In Proceedings of the Instructional Technology and Education of the Deaf: Supporting Learners, K- College: An International Symposium, Rochester, NY, USA, 25–27 June 2001.
3. Lopéz, D.; Cruz, J.-L.; Sánchez, F.; Fernández, A. A take-home exam to assess professional skills. In Proceedings of the 41st ASEE/IEEE Frontiers in Education Conference, Rapid City, SD, USA, 12–15 October 2011.
4. Rich, R. An experimental study of differences in study habits and long-term retention rates between take-home and in-class examination. *Int. J. Univ. Teach. Fac. Dev.* **2011**, *2*, 123–129.
5. Biggs, J. *Teaching for Quality Learning at University*; Oxford University Press: Oxford, UK, 1999.
6. Giammarco, E. An Assessment of Learning Via Comparison of Take-Home Exams Versus In-Class Exams Given as Part of Introductory Psychology Classes at the Collegiate Level. Ph.D. Thesis, Capella University, Minneapolis, MN, USA, 2011.
7. Andersson, L.; Krathwhol, D.; Airasian, P.; Cruikshank, K.; Mayer, R.; Pintrich, P.; Wittrock, M.C. *A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*; Pearson: London, UK, 2001.
8. Andrada, G.; Linden, K. Effects of two testing conditions on classroom achievement: Traditional in-class versus experimental take-home conditions. In Proceedings of the Annual Meeting of the American Educational Research Association, Atlanta, GA, USA, 12–16 April 1993.
9. Bloom, B. *Taxonomy of Educational Objectives. The Classification of Educational Goals, Handbook 1: Cognitive Domain*; David McKay: New York, NY, USA, 1956.
10. Svoboda, W. A case for out-of-class exams. *Clear. House J. Educ. Strateg. Issues Ideas* **1971**, *46*, 231–233. [CrossRef]
11. Bredon, G. Take-home tests in economics. *Econ. Anal. Policy* **2003**, *33*, 52–60. [CrossRef]

12. Zoller, U. Alternative Assessment as (critical) means of facilitating HOCS-promotion teaching and learning in chemistry education. *Chem. Educ. Res. Pract. Eur.* **2001**, *2*, 9–17. [CrossRef]
13. Carrier, D. Legislation as a stimulus to innovation. *High. Educ. Manag.* **1990**, *2*, 88–98.
14. Aggarwal, A. A Guide to eCourse Management: The stakeholders' perspectives. In *Web-Based Education: Learning from Experience*; Aggarwal, A., Ed.; Idea Group Publishing: Baltimore, MD, USA, 2003; pp. 1–23.
15. Hall, L. Take-home tests: Educational fast food for the new Millennium? *J. Manag. Organ.* **2001**, *7*, 50–57. [CrossRef]
16. Bell, A.; Egan, D. The Case for a Generic Academic Skills Unit. Available online: http://www.qualityresearchinternational.com/esecttools/esectpubs/belleganunit.pdf (accessed on 15 January 2019).
17. Haynie, W. Effects of take-home and in-class tests on delayed retention learning acquired via individualized, self-paced instructional texts. *J. Ind. Teach. Educ.* **1991**, *28*, 52–63.
18. Durning, S.; Dong, T.; Ratcliffe, T.; Schuwirth, L.; Artino, A., Jr.; Boulet, J.; Eva, K. Comparing open-book and closed-book examinations: A systematic review. *Acad. Med.* **2016**, *91*, 583–599. [CrossRef]
19. Ramsey, P.; Carline, J.; Inui, T.; Larsson, E.; Wenrich, M. Predictive validity of certification. *Annu. Intern. Med.* **1989**, *110*, 719–726. [CrossRef]
20. Gough, D. Weight of evidence: A framework for the appraisal of the quality and relevance of evidence. *Appl. Pract. Based Res.* **2007**, *22*, 213–228. [CrossRef]
21. Bearman, M.; Smith, C.; Carbone, A.; Slade, S.; Baik, C.; Hughes-Warrington, M.; Neuman, D. Systematic review methodology in higher education. *High. Educ. Res. Dev.* **2012**, *31*, 625–640. [CrossRef]
22. Freedman, A. The take-home examination. *Peabody J. Educ.* **1968**, *45*, 343–347. [CrossRef]
23. Marsh, R. Should we discontinue classroom tests? An experimental study. *High Sch. J.* **1980**, *63*, 288–292.
24. Weber, L.; McBee, J.; Krebs, J. Take home tests: An experimental study. *Res. High. Educ.* **1983**, *18*, 473–483. [CrossRef]
25. Marsh, R. A comparison of take-home versus in-class exams. *J. Educ. Res.* **1984**, *78*, 111–113. [CrossRef]
26. Grzelkowski, K. A Journey Toward Humanistic Testing. *Teach. Sociol.* **1987**, *15*, 27–32. [CrossRef]
27. Zoller, U.; Ben-Chaim, D. Interaction between examination type, anxiety state and academic achievement in college science; an action-oriented research. *J. Res. Sci. Teach.* **1989**, *26*, 65–77. [CrossRef]
28. Murray, J. Better testing for better learning. *Coll. Test.* **1990**, *38*, 148–152. [CrossRef]
29. Fernald, P.; Webster, S. The merits of the take-home, closed book exam. *J. Hum. Educ. Dev.* **1991**, *29*, 130–142. [CrossRef]
30. Ansell, H. Learning partners in an engineering class. In Proceedings of the 1996 ASEE Annual Conference, Washington, DC, USA, 23–26 June 1996.
31. Norcini, J.; Lipner, R.; Downing, S. How meaningful are scores on a take-home recertification examination. *Acad. Med.* **1996**, *71*, 71–73. [CrossRef] [PubMed]
32. Haynie, J. Effects of take-home tests and study questions on retention learning in technology education. *J. Technol. Learn.* **2003**, *14*, 6–18. [CrossRef]
33. Tsaparlis, G.; Zoller, U. Evaluation of higher vs. lower-order cognitive skills-type examination in chemistry; implications for in-class assessment and examination. *Univ. Chem. Educ.* **2003**, *7*, 50–57.
34. Giordano, C.; Subhiyah, R.; Hess, B. An analysis of item exposure and item parameter drift on a take-home recertification exam. In Proceedings of the Annual Meeting of the American Educational Research Association, Montreal, QC, Canada, 11–15 April 2005.
35. Moore, R.; Jensen, P. Do open-book exams impede long-term learning in introductory biology courses? *J. Coll. Sci. Teach.* **2007**, *36*, 46–49.
36. Frein, S. Comparing In-class and Out-of-Class Computer-based test to Traditional Paper-and-Pencil tests in Introductory Psychology Courses. *Teach. Psychol.* **2011**, *38*, 282–287. [CrossRef]
37. Marcus, J. Success at any cost? There may be a high price to pay. *Times Higher Education London*, 27 September 2012; 22–25.
38. Tao, J.; Li, Z. A Case Study on Computerized Take-Home Testing: Benefits and Pitfalls. *Int. J. Tech. Teach. Learn.* **2012**, *8*, 33–43.
39. Hagström, L.; Scheja, M. Using meta-reflection to improve learning and throughput: Redesigning assessment producers in a political science course on power. *Assess. Eval. High. Educ.* **2014**, *39*, 242–253. [CrossRef]
40. Rich, J.; Colon, A.; Mines, D.; Jivers, K. Creating learner-centered assessment strategies or promoting greater student retention and class participation. *Front. Psychol.* **2014**, *5*, 1–3. [CrossRef]

41. Sample, S.; Bleedorn, J.; Schaefer, S.; Mikla, A.; Olsen, C.; Muir, P. Students' perception of case-based continuous assessment and multiple-choice assessment in a small animal surgery course for veterinary medical students. *Vet. Surg.* **2014**, *43*, 388–399. [CrossRef]
42. Johnson, C.; Green, K.; Galbraith, B.; Anelli, C. Assessing and refining group take-home exams as authentic, effective learning experiences. *Res. Teach.* **2015**, *44*, 61–71.
43. Downes, M. University scandal, reputation and governance. *Int. J. Educ. Integr.* **2017**, *13*, 8. [CrossRef]
44. D'Souza, K.; Siegfeldt, D. A conceptual framework for detecting cheating in online and take-home exams. *Decis. Sci. J. Innov. Educ.* **2017**, *15*, 370–391. [CrossRef]
45. Lancaster, T.; Clarke, R. Rethinking assessment by examination in the age of contract cheating. In Proceedings of the Plagiarism Across Europe Beyond, Brno, Czech Republic, 24–26 May 2017; ENAI: Brno, Czech Republic; pp. 215–228.
46. Foley, D. Instructional potential of teacher-made tests. *Teach. Psychol.* **1981**, *8*, 243–244. [CrossRef]
47. Kaplan, H.; Sadock, B. Learning theory. In *Synopsis of Psychiatry: Behavioral Science/Clinical Psychiatry*, 8th ed.; Kaplan, H., Sadock, B., Eds.; Williams &Wilkins: Philadelphia, PA, USA, 2000; pp. 148–154.
48. Berett, D. Harvard Cheating Scandal Points Out Ambiguities of Collaboration. *Chron. High. Educ.* **2012**, *59*, 7.
49. Pennington, B. Cheating scandal dulls pride in Athletics at Harvard. *The New York Times*, 18 September 2012; B11.
50. Conlin, M. Commentary: Cheating or postmodern learning? *Bus. Week* **2007**, *4043*, 42.
51. Young, J. Cheating incident involving 34 students at Duke is Business School's biggest ever. *Chronicle of Higher Education*, 30 April 2007.
52. Coughlin, E. West Point to Review 823 Exam Papers. *Chronicle of Higher Education*, 31 May 1976; 12.
53. Vassar, R. Take-home exams are not the solution. *The Stanford Daily*, 8 October 2012.
54. Agarwal, P.; Karpicke, J.; Kang, S.; Roediger, H., III; McDermott, K. Examining the Testing Effect with Open- and Closed-Book Tests. *Appl. Cognit. Psychol.* **2008**, *22*, 861–876. [CrossRef]
55. Ebel, R. *Essentials of Educational Measurement*; Prentice Hall: Englewood Cliffs, NJ, USA, 1972.