Cognitive Comparability of Grade 12 Lebanese National Examinations in Chemistry from 2001 to 2012

Zeina Hajo and Samar Zeitoun(*)

Abstract

The purpose of this study is to investigate whether the cognitive demands in Chemistry National Exams (CNE) in Lebanon for Grade 12 have been maintained over time. This study is carried out to evaluate the appropriateness of using CNE as parallel forms having same overall difficulty -as they are used in Lebanon. The test items of 23 CNE at Grade 12 Life Science section from 2001 to 2012 (1st sessions and 2nd sessions) were analyzed and categorized using a four-level taxonomy developed by Webb and his colleagues (Webb et al., 2005) to classify science objectives and science test items. Results show that the cognitive demands of CNE for Grade 12 Life Science section are not maintained neither from 1st session (normal session) to 2nd session (exceptional session) nor from year to year.

Introduction

National examinations gain a particular interest in Lebanon for students, schools and policymakers. During the examination period, the national exams draw a great deal of attention inside and outside the media. While students and the public are merely interested in the fairness of the examinations, policymakers are focused on other educational issues especially in times of major reforms in education. Every year, when examinees terminate their national exams they begin comparing the test difficulty in the current session with the test difficulty in the previous sessions in every subject (e.g., Chemistry, Physics, and Math). Many teachers and students feel that there is variation in the difficulty (cognitive demands) of the LNE from first session to second session and from year to year. Therefore, they consider the examinations to be unfair.

(*) Lebanese University, Faculty of Education, Beirut, Lebanon.

Indeed, the consecutive sessions of the LNE in a particular subject (e.g. Chemistry) are used as parallel versions of an exam. This is based on the assumption that all parallel versions are equivalent in difficulty. The equivalence in difficulty is one of three conditions (will be explained later) that allows examination boards to claim comparability of exams and fairness among examinees of different sessions (AERA et al., 1999; AQA, 2005; Baird, 2007; Downing, 2006; Newton, 2007). The problem is that the use of the exams as parallel versions and the claim that they are equivalent in difficulty is not evidence based. The aim of this paper is to:

* Investigate whether the cognitive demands in Chemistry National Exams (CNE) in Lebanon for Grade 12 have been maintained over time.
* Evaluate the appropriateness of using CNE as parallel versions having same overall difficulty - as they are used in Lebanon.

Review of relevant literature

After the last educational reform in Lebanon in 1997, as before the reform, the Lebanese National Examinations (LNE) are always developed for certification purpose (to certify students' achievement of the intermediate or the secondary educational stage). They are criterion-referenced exams where government standards (a set of criteria) are defined and the students' performance is judged by referring to the predetermined standards (Dunn et al., 2004; Popham, 1999). So the student performance in the national exam in a particular subject (e. g., Chemistry) reflects the extent to which the course contents represented by the exam are being met by the student. The pass score in the LNE continue to be 50%. Examinees with scores above 50% are considered to have demonstrated an accepted level of ability and therefore pass the examination or vice versa.

National examinations - including LNE- are developed to meet government standards and to maintain those standards over different exams (Kellaghan et al., 2009; Newton, 2007; SCAA, 1996).The maintenance of standards in national examinations is a matter of national interest (SCAA, 1996; Kruger, 2010). However, its social and political significance differs from one country to another (Wolf, 2000). In England, for example, where several forms of comparability (between examining boards, over time, and across subjects) are applied, maintenance of standards is a major concern for stakeholders (Newton, 2007).

It was described by Murphy and Broadfoot (1995) as «the English disease» (p. 54).

To be clear, it is worth at this point to share a common understanding of what are «Standards». Several meanings are attached to the word «Standards» but the main dictionary definitions are two. One is that a standard is «something set up and established by authority *as a rule for the measure of quantity, weight, extent, value or quality»* while the other one define standards as «something established by authority, custom or general consent *as a model or example*» (Wolf, 2000, p. 14). This study adopts the second definition of standards as an example and ideal that should be met by the national exams, not standards as tools for measuring. Therefore, in this research study, the national examination standards refer to (SCAA, 1996):

- the demands of syllabi and their assessment arrangements (*the examination demands*) (e.g., the cognitive demands of Grade 12 Chemistry National Exams), and
- the level of performance required of candidates to gain particular grades (*the grade demands*) (e.g., the marking schemes of Chemistry National Exams).

This study focuses more on standards as examination demands since it aims to investigate whether the cognitive demands in Chemistry National Exams (CNE) in Lebanon for Grade 12 have been maintained over time. But why are we talking about maintenance of examination standards? Do national examination standards vary?

Newton (2007) explains that standards in national examinations might vary:

- over time (e.g. it is easier to pass a particular exam in 2011 than in 2012),
- between examination boards -when two or more boards are responsible to set exams (e.g. it is easier to pass a particular exam with board A rather than board B), and
- across subjects- when optional subjects are offered (e.g. it is easier to pass the Chemistry exam rather than the Physics exam).

So, maintenance of examination standards or *comparability* is the application of the same standards (examination demands and grade demands) across different examinations.

In Lebanon, where a single board is responsible to set national examinations and all subject matters tested in the LNE are compulsory, the only type of comparability to investigate is comparability over time.

When dealing with comparability over time the selected exams to study are parallel versions. The consistency in content and cognitive demands is a condition in developing parallel exam forms (Jones et al., 2006). Parallel versions are distinct forms of an exam having same level of difficulty and have been constructed to represent the same content and to satisfy the same test specification; they differ only in terms of the test items posed (AERA *et al.*, 1999, Newton, 2007). When the examination system allow failing students to retake the test, such as the Lebanese national examination system, the tests are referred to as concurrent parallel form (Jones et al., 2006). Concurrent parallel forms are equivalent forms of an exam with same level of difficulty (cognitive demands) from year to year.

An error of measurement occurs when there is difference in difficulty between parallel forms used (AERA et al., 1999). If this year test version is more difficult than the last year version, students who take the more difficult version may get a lower grade than those who take the easier version. So a certain score (raw mark) earned this year is not equivalent to the same score earned last year. Consequently the same score awarded in successive years does not represent the same performance.

However, the difficulty level of parallel versions is not the only factor that affects the equivalence of scores from session to session, but it is the first condition for comparability. Two other necessary conditions for comparability are: developing marking schemes that give fair rewards over time (Robinson, 2007) and adjusting scores every year (Newton, 2007). Comparability, as defined earlier, is the application of the same standards (examination demands and grade demands) across different examinations. While parallel versions with same level of difficulty assure the consistency of examination demands from session to sessions, developing reliable marking scheme and adjusting scores guarantee the application of the same grade demands across consecutive years. Indeed, the last two conditions relevant to the grade demands will not be detailed since they are beyond the scope of this study.

Comparability, fairness, validity and all other principles relevant to test credibility are promoted through following a well designed examination process

and an effective test development (AERA et al., 1999; AQA, 2005; Downing, 2006). Downing (2006) discussed twelve steps in test development that, if adequately followed, help to insure high credibility of the test developed. Comparability as maintenance of examination demands are mainly related to the second step explained by Downing - establishing test specification. The main condition to develop parallel forms of an exam is to construct them based on the same test specification (AERA et al., 1999; Newton, 2007). «The test specifications form an exact sampling plan for the content domain» (Downing, 2006, p. 9).

The test specifications should be established based on clear method that may combine empirical and rational/judgmental methods (Downing, 2006). The Standards (AERA et al., 1999) emphasize documenting the test specifications with the rationale and process by which they were developed.

> Standard 3.3 of the Standards explains what the test specifications should define:
> the content of the test, the proposed number of items, the item format, the desired psychometric properties of the items, and the item and section arrangement. They should also specify the amount of time for testing, directions to the test takers, procedures to be used for test administration and scoring, and other relevant information. (AERA. *et al.*, 1999, p. 43)

Linn (2006) mentions that in addition to the content coverage, test specifications should identify the cognitive processes used by the examinees and the extent to which these processes cover the cognitive demands of the domain measured. Cognitive demand is the amount of intellectual activity required to perform a task. Downing (2006) distinguishes between test specifications that offer an operational definition of all test characteristics (as detailed in Standards 3.3 above) and test blueprints which form part of test specifications. To form a test blueprint, both test content and cognitive processes are used together. Test blueprint is a two-dimensional (content by cognitive demands) table that generates a relationship between the content domains (or objectives) that should be assessed and the items that appear on the test (Bridge.et al., 2003; Downing, 2006; Linn, 2006; Popham, 2005).

Parallel versions of an exam constructed based on the same test

specification that includes a clear test blueprint must be equivalent in content (content comparability) and in cognitive demands (cognitive comparability). This is the best way to assure comparability as maintenance of examination demands across different examinations.

A framework to analyze the Lebanese Chemistry National Exams

The purpose of this research is to investigate the cognitive comparability of Grade 12 CNE in an attempt to find out whether these exams are parallel versions having same level of difficulty. To accomplish this task, the cognitive complexity of each test item in the exam must be determined. The cognitive complexity refers to the cognitive demand associated with an item. The rationale for classifying items by their level of complexity is to focus on the expectations of the item, not the ability of the student. The item's demands-what the item requires the student to do (recall, understand, analyze,) -are made with the assumption that the student is familiar with the basic concepts of the task.

Several studies have looked at the cognitive demand of tasks as they are written and implemented in the classroom. Looking at the cognitive demand is important because it helps identify the level of thinking processes that occur in a student's mind. There are several ways to categorize tasks based on cognitive demand: Bloom's revised taxonomy (Anderson & Krathwol, 2001), Marzano (2001), Merill's performance content matrix (1994), PISA (OECD, 1999), Porter &Smith (2001)...

A taxonomy that defines different cognitive levels was needed to classify the test items. Although the famous Bloom's taxonomy was a possible instrument to chose, a four-level taxonomy developed by Webb and his colleagues (Webb et al., 2005 a, b) was adopted as a framework for this study for three reasons:

(a) Verbs are not the only identifier in determining item complexity but, instead, each level depends on how deeply students understand the content in order to answer the test item.

(b) The «Depth-of-Knowledge levels» are nominative levels particularly designed to specific content areas such as science. The «Science Depth-of-Knowledge levels» are used to classify objectives and science test items according to their cognitive complexity.

(c) The clear examples provided by Hess (2010) to distinguish between the four

cognitive levels and to illustrate how science test items could be classified accordingly (see table 1).

The four general cognitive levels explained by Webb (2002) include the recall and reproduction (Level 1), the skills and concepts (Level 2), the strategic thinking (Level 3), and the extended thinking (Level 4). The «Science Depth-of-Knowledge levels», which consists of four cognitive levels for science, developed by Webb and his colleagues in 2005 is adopted for use in this study. The following description of the «Science Depth-of-Knowledge levels» is based on Webb et al. (2005b) and Hess (2010).

The recall and reproduction level (Level 1) only requires students to demonstrate a rote response, use a well-known formula, follow a simple procedure typically involves only one-step. To answer a level 1 item the student either knows or does not know the answer. The item does not need to be figured out or solved.

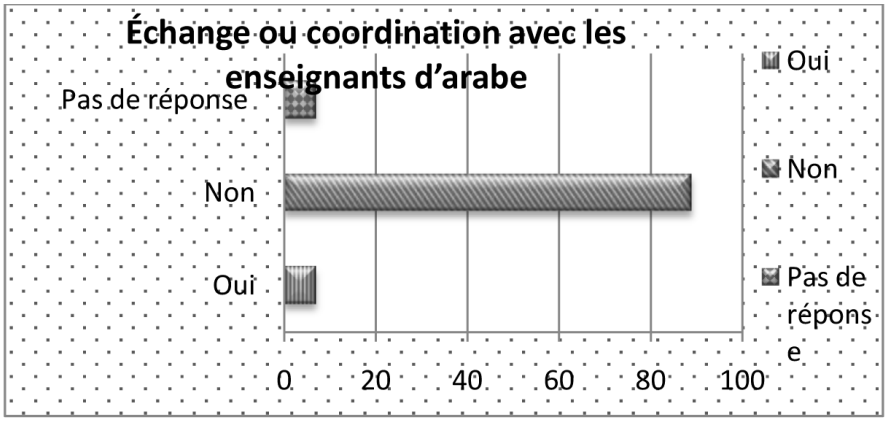
The skills and concepts level (Level 2) includes the engagement of some mental processing beyond recalling or reproducing a response. An item at this level requires students to make some decisions as to how to approach a question or problem. These actions generally imply more than one step.

The strategic thinking level (Level 3) includes items that require reasoning, planning, using evidence, and thinking at a higher level than the previous two levels. In most instances, items at this level require students to explain their thinking, generate conclusions from observations, use concepts to explain phenomena, and solve non-routine problems.

At the extended thinking level (Level 4) high cognitive demands and complex work are required. Students are asked to make connections within and between subject matters (e.g. Chemistry and Physics). This level includes items that require complex reasoning, experimental design and planning, and thinking for an extended period of time. Table 1 represents clear examples for each of the cognitive levels in science offered by Hess (2010).

The test items of the analyzed CNE in this study were assigned to the four cognitive levels of Webb's taxonomy based on the description of the «Science Depth-of-Knowledge levels» and the examples mentioned in table 1. Consequently, items of CNE that require recalling names, formulas, equations, definitions and properties of specific chemical reagents, and items that require

direct applications of formulas were considered at the recall and reproduction level (level 1). Example of CNE items assigned to level 1 are: *«name the type of the catalysis carried out in the presence of a platinum wire», «indicate why the taken volume is poured in ice-cold distilled water before titration»*, and *«complete the equation of the oxidation reaction of ethanal by Tollens' reagent»*. Although the equation in the last example looks complicated but the item was assigned to level 1 as it only requires recall knowledge.



Items of CNE classified at the skills and concepts level(level 2) included giving systematic names based on structural formulas or vice versa, plotting a graph that shows relations, making calculations that require more than one step (whereas simple substitutions in relations were considered at level 1). Example of CNE items at level 2 are: *«write the condensed structural formula of alcohol B, knowing that its systematic name is 3-methyl-1-butanol» and «Plot the curve* $n_{(ester)} = f(t)$ *in the interval of time [0; 30 min]»*. In addition, items such as *«write the equation of the occurred reaction»* were considered at level 1 when chemical formulas are given and the student has only to recall the obtained products and at level 2 if the student has to figure out the chemical formulas. Items like *«write the formula of the compound (F)»* were assigned to level 1 when the student has to recall the formula and assigned to level 2 when the student has to use the law of conservation of matter to find the formula.

Items of CNE classified at the strategic thinking level (level 3) might ask the student to establish or show a relationship (e.g., *show that the amount of mater of ester in volume* V_{tot}*, at any instant t, is given by the relation* $n\ (ester)_t = 0.383 - V_E$*), give arguments to validate or invalidate a statement (e.g., Justify or correct: The* pK_a *of the conjugate pair NH4 / NH3 is given by the following relationp* $H = pK_a + log\ (CaVa\text{-}CaVae)/CbVb)$, draw conclusions (e.g., *justify the term buffer based on properties of solutions*), interpret from a complex graph (e.g., *justify, based on the above table and graph, whether the reaction in each of the mixtures A and B, has finished at the instant t = 25 min*). Concerning the extended thinking level (level 4), no items of the analyzed CNE were assigned to this level.

One additional aspect needs to be mentioned in connection with routine and non-routine test items. Anderson (2005) clearly indicates that the questions the learners have experienced previously would be placed in a lower cognitive level than what is should be actually placed as they become routine problems. In this study the researchers were aware of this point. Therefore they started from the 1st session in 2001 to reach the 2nd session in 2012 following the chronological order and then paid attention to any recycled questions in later examinations as students are frequently coached for the final exams by working through previous exam papers. For example, the same test item «*by specifying its remarkable points, trace the shape of the curve representing the variation of pH of the content of the beaker versus the added volume* V_a *of the acid*» appeared in

three sessions, 1st session 2004, 1st session 2006, and 2nd session 2008. While this item was assigned to cognitive Level 3 in the first two sessions, it was assigned to level 2 in session 2008.

Procedure

In this study the test items of 23 CNE at Grade 12 Life Science section from 2001 to 2012 were analyzed and categorized using the framework developed by Webb et al. (2005 b), «the Science Depth-of-Knowledge levels», and the table of examples offered by Hess (2010) as explained earlier. The 23 exams are 12 normal sessions (1st sessions) and 11exceptional sessions (2nd sessions). The 2004 (2nd session) was exempted from this study due to the unavailability of the detailed marking scheme.

In Lebanon, National Examinations are conducted in two sessions every year: the normal session (1st session) and the exceptional session (2nd session). The exceptional session is a second chance for students who fail the normal session. The exams in both sessions are identical in format. Concerning the Grade 12 CNE, each exam consists of three independent compulsory exercises. Each exercise includes several test items. The exam is marked on a total of 20 but has a weight 4 in the LNE. Thus, the total mark attributed to CNE is 80 points.

In general, a test item is «a single, often decontextualized test question or problem» (NCTM, 1995, p. 88). More specifically, this paper adopts the definition of a test item as being any part of the test that requires a response from the student which entitles him/her to a part of the grade. A test item may take one of the two following forms: a question that requires an answer. For example, *«How does this rate vary with time?»* or an imperative sentence, such as *«Give the condensed structural formula of the carboxylic acid A»*. In the case of many components required in one sentence, it is considered to stand for more than one test item. For example, *«Indicate the role of heating and that of sulphuric acid»* is counted for two items, because it stands for *«indicate the role of heating» and «indicate the role of sulphuric acid.»*

It is important to mention that in data analysis the point values of the original assessment items that appear in the exams are taken into consideration. The original item of two-point value, for example, was considered as though it were two identical one-point items based on the recommendation of Webb et al. (2005b). Therefore, the items found in the analyzed exams are termed *original*

test items, and the items considered in this study (each with one point value) are termed *assessment items*. While the total score of Chemistry in the LNE is 80 points, it is assumed that each CNE consists of 80 assessment items.

The authors conducted the analysis and categorization of the test items based on their own judgments. The two researchers are experts in the field since both have a Masters degree in Chemistry and taught Grade 12 Chemistry for a certain number of years. To increase validity and reliability of findings, the two researchers met to discuss the «Science Depth-of-Knowledge levels» and the examples in table 1 in detail to identify the main characteristics of each level in order to reach a common understanding. After the meeting, they chose the 1st session 2001and determined jointly the cognitive complexity of all its items by classifying each item as belonging to Level 1, Level 2, Level 3, or Level 4 of the framework. This was done to ensure consistency of analysis. Then, the two authors categorized independently the test items of the other exams and met after analyzing each exam to reconcile discrepancy in categorization and reach consistency. They started from the 1st session in 2001 to reach the 2nd session in 2012 following the chronological order. Subsequently, the frequencies and percentages of assessment items in each cognitive level for each of the 23 analyzed CNE were computed. The results of this categorization are presented in tables 2-6.

Results

The percentage distribution of the cognitive levels of the CNE from 2001 to 2012 (1st session and 2nd session) is presented in table 2 and figure 1. This figure shows that all sessions, include items at the recall and reproduction level (Level 1), the skills and concepts level (Level 2), and the strategic thinking level (Level 3). The exception is first session 2001 which is formed of items at levels 1 and 2 only. This could be attributed to the fact that this was the first national exam administered after reform. To make the exam easy to pass, the examination board did not include items at level 3. Table 2 and figure 1 show also that CNE do not include any item at the extended thinking level (Level 4). This was expected as Level 4 requires complex reasoning, planning, developing, and thinking most likely over an extended period of time.

Objectifs	Activités en langue arabe	Activités en langue française
-Anticipation et formulation de l'histoire en arabe. **-Formulation de l'histoire en langue 2.** **-Identification des mots (lecture guidée du texte) en L1 et L2.**	**Phase 1** Activité d'expression orale + d'anticipation avant la lecture +écriture au tableau des prévisions des élèves. Anticipation du contenu du document. Emission d'hypothèses de sens. Le maitre note au tableau les propositions des élèves **Phase 2** Activité d'identification des mots écrits en arabe.	**Phase 1** Formulation en français de l'histoire déjà lu en arabe. **Phase 2** Activité d'identification des mots écrits en français.

Figure 1 Percentage distribution of the cognitive levels for the CNE from 2001 to 2012 (1st session and 2nd session) (N= normal session = 1st session) and E= exceptional session = 2nd session)

Table 1 shows that in average 53.75% of the assessment items of the analyzed CNE are at level 2 while the average of items at level 1 is 23.04% and the average of items at level 3 is 23%. This result indicates that in general the Lebanese CNE highly emphasize the skills and concepts level since, in average, more than half of the items posed over 12 years are at this cognitive level. It also indicates that, in general, the recall and reproduction level and the strategic thinking level are equally emphasized in the CNE.

Table 2 Percentage distribution of the cognitive levels for the CNE from 2001 to 2012 (1st session and 2nd session)

Objectifs	Activités en langue arabe	Activités en langue française
-Lien sujet –verbe	-Exercice d'appariement Lier les personnages – aux actions (sujet et verbe).	Repérage de la ponctuation dans le texte. Valeur des signes de ponctuation. Exercice.
-Les valeurs des signes de ponctuation.	-Exercice sur la ponctuation. Transfert des acquis de L2 vers L1.	Exercice d'appariement Lien personnage –actions faire le même exercice de L1 en L2. Transfert des acquis de L1 vers L2.
-Structure du récit.	-Structure du récit : Exercice de mise en ordre des éléments du récit.	Structure du récit : même exercice en L2 Consolidation des acquis et lecture globale des phrases.
-Les marqueurs temporels.	-Exercice lacunaire Les marqueurs de temps dans le récit en L1.	-Exercice lacunaire Les marqueurs de temps dans le récit en L2
-Lien entre graphème et phonème.	-Exercice d'identification : Rappel lien graphème phonème le ا أ	-Exercice d'identification Rappel graphème phonème Le i

Indeed, the average percentages (in the last row of table 1) do not give any idea about the cognitive comparability of CNE. However, table 2 and figure 1 show the variation of the cognitive complexity of items over the 23 analyzed tests. The percentage of items at level 1 varies between 5% (2nd session 2011) and 43.75% (1st session 2001) with a difference of 38.75. The percentage of items at level 2 varies between 35% (1st session 2007 and 1st session 2008) and 81.25% (1st session 2011) with a difference of 46.25. At level 3, the percentage of items changes from 0% (1st session 2001) to 41.25% (1st session 2004) with a difference of 41.25. The huge difference between the highest and lowest percentage of items at each cognitive level (38.75 for level 1, 46.25 for level 2, and 41.25 for level 3) from session to session over 12 years indicates that the cognitive demands were not maintained in CNE.

To reach accurate results about the variation of cognitive demands for the same session (1st session and 2nd session) from year to year, the percentage

distribution of the cognitive levels in the items of the 1st sessions of CNE from 2001 to 2012 is presented in figure 2, while the percentage distribution of the cognitive levels in the items of the 2nd sessions of CNE from 2001 to 2012 is presented in figure 3. The inconsistency in cognitive demands is clearly shown in the two figures as the percentage distribution for each of the three levels varies significantly from year to year for the 1st sessions (figure 2) and for the 2nd session (figure 3)

Objectifs	Activités en langue arabe	Activités en langue française
-Lecture de courts textes documentaires.	-Identification des mots pour lire un court texte documentaire sur les caractéristiques de la souris	- Identification des mots pour lire un court texte documentaire sur les caractéristiques du lion.
-Réalisation des maquettes à partir des fiches techniques	-Lire en groupe une fiche technique pour réaliser une activité pratique : maquette d'une souris à partir des consignes écrites.	-Suite à une lecture en plénière d'une fiche technique pour la fabrication d'une maquette de lion miniature, les enfants réalisent l'activité.

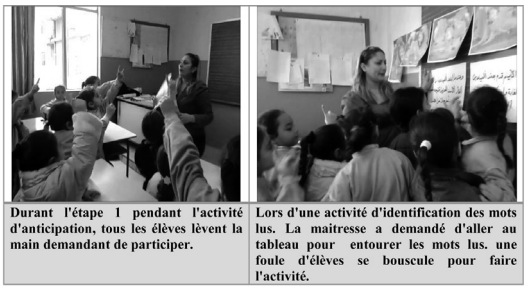
Figure 2 Percentage distribution of the cognitive levels in the items of the 1st sessions of CNE from 2001 to 2012

Objectifs	Lectures possibles autour du thème en arabe	Lectures possibles autour du thème en français
-Développer les capacités de compréhension. **-Bâtir une culture.**	كليلة ودمنة الاسد والثور. الاسد والارنب. قصة الفأر والحمامة الفأر الطماع	-"Hibou blanc et souris bleue" de jean Joubert. -"Le lion magicien" de Catherine Missonnier. -"Souricette" de Myriam Deru. -"Frédérique" de Léo Lionni. -"L'arbre sans fin" de Claude Ponti

Figure 3 Percentage distribution of the cognitive levels in the items of the 2nd sessions of CNE from 2001 to 2012

By comparing the data from both figures 2 and, 3 it is clear that the highest percentage of items was for level 2 in both first and second sessions. This result indicates that CNE give emphasis to the skills and concepts level (level 2) in first sessions as well as in second sessions. However the figures show that there is no regular pattern for level 1 and level 3.

The cognitive demands of the first sessions and the second session in the same year were also compared. Figures 4, 5 and 6 present the percentage of items in the first session and the second session of CNE from 2001 to 2012 for the cognitive levels 1, 2 and 3 respectively. Figures 4, 5 and 6 show huge discrepancies in the percentages of test items for the three cognitive levels across the first and the second sessions for the same year.



Durant l'étape 1 pendant l'activité d'anticipation, tous les élèves lèvent la main demandant de participer.

Lors d'une activité d'identification des mots lus. La maitresse a demandé d'aller au tableau pour entourer les mots lus. une foule d'élèves se bouscule pour faire l'activité.

Figure 4 Percentage of items at the first session and the second session of CNE from 2001 to 2012 at the cognitive levels 1

	Nombre d'Interventions des élèves Cours de Français
Séance 1 avant	163
Séance 2 pendant l'expérience	254

Figure 5 Percentage of items at the first session and the second session of CNE from 2001 to 2012 at the cognitive levels 2

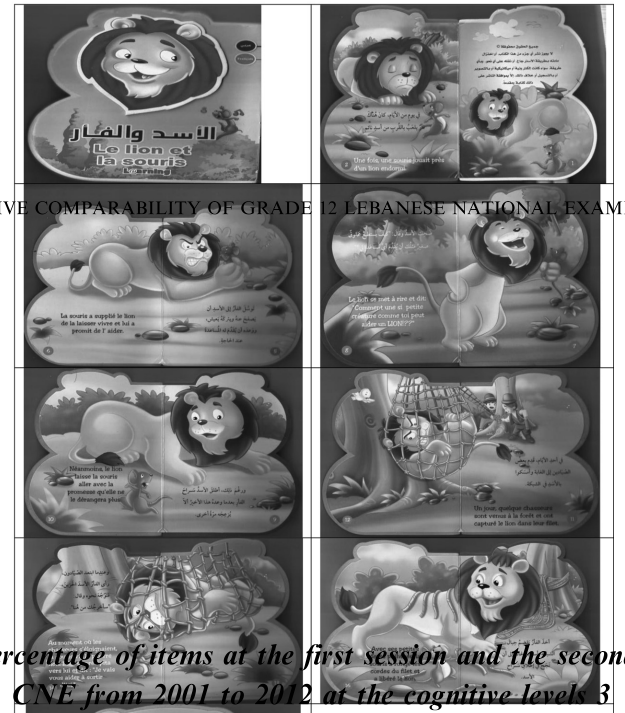


Figure 6 Percentage of items at the first session and the second session of CNE from 2001 to 2012 at the cognitive levels 3

This result was confirmed when the correlation between the first and the second sessions for each cognitive level (level 1, 2 and 3) was computed over eleven years (session 2004 was excluded).This correlation was calculated for each level using Pearson Product-Moment coefficient under Microsoft Excel by correlating data in two tables (one table for the 1st sessions and the other table for the 2nd sessions) cell by cell where each cell corresponds to a year. The value of the coefficient varies between +1 and-1 inclusive, where 1 is total positive correlation, 0 is no correlation, and-1 is total negative correlation. The obtained correlation coefficients are presented in table 3. This table shows a very low correlation between the normal session and the exceptional session for each of the three cognitive levels since the three computed coefficient are very close to zero.

Table 3 Person correlation coefficient betwe CNE fro vel

page 2 في يوم من الأيام، كان هناك فأر يلعب بالقرب من أسد نائم. Une fois, une souris jouait près d'un lion endormi.
Page 3-4 فجأة إستيقظ الأسد غاضباً، فأمسك الفأر في قبضته و هدده أن يأكله. Tout à coup, le lion se réveilla, prit la souris dans son poing et l'a menacée de la manger.
Page 5-6 توسل الفار إلى الأسد أن يصفح عنه ويتركه يعيش، ووعده أن يقدم له المساعدة عند الحاجة. La souris a supplié le lion de la laisser vivre et lui a promis de l'aider.
page 8 ضحك الأسد وقال: "كيف يستطيع مخلوق صغير مثلك أن يقدم أي مساعدة لي؟" Le lion se met à rire et dit : « Comment une si petite créature comme toi peut aider un LION !?? »
page 9-10 ورغم ذلك، أطلق الأسد سراح الفأر بعدما وعده هذا الأخير الا يزعجه مرة أخرى. Néanmoins, le lion laisse la souris aller avec la promesse qu'elle ne le dérangera plus.
page 11 في أحد الأيام، قدم بعض الصيادين إلى الغابة و أمسكو بالأسد في الشبكة. Un jour, quelques chasseurs sont venus à la foret et ont capturé le lion dans leur filet.
page 14 و عندما إبتعد الصيادون، رأى الفأر الأسد الحزين، فتوجه نحوه وقال: "سأخرجك من هنا". Au moment où les chasseurs s'éloignent, la souris vit le pauvre lion. Elle se précipita vers lui et dit : « je vais vous aider à sortir d'ici. »
page 15-16 أخذ الفار يقدم حبال الشبكة بأسنانه إلى أن نجح في إطلاق سراح الأسد . Avec ses petites dents pointues, la souris a rongé les cordes du filet et a libéré le lion.
page 17 ومنذ ذلك الحين، أصبح الأسد و الفأر صديقين حميمين. Dès ce jour-là, le lion et la souris sont devenues des meilleurs amis.

Discussion

The purpose of this research was to investigate the cognitive comparability of Grade 12 CNE in an attempt to find out whether or not theses exams are parallel versions having same level of difficulty (Kruger, 2012).

The results of this study show that the cognitive demands of CNE for grade 12 Life Science section are not maintained from session to session. The cognitive complexity of the exams varies significantly from year to year and from the first session to the second session in the same year. Based on these results, the successive sessions of grade 12 CNE Life Science section are not parallel versions since they are not at the same difficulty level. The significant variation in cognitive demands between the analyzed exams poses serious questions about fairness of examination and the use of CNE as parallel versions in Lebanon.

While examination boards, in general, do their best to set parallel versions with same level of difficulty year to year, the board of CNE in Lebanon does not take the cognitive demands in consideration when the exams were developed. This is due to the brief test specification based on which the CNE is constructed and the absence of test blueprint that map the cognitive complexity of the items.

The test specification of the Grade 12 CNE Life science section -named the General Instructions- is presented in table 4. This test specification is very few and not specific enough. It does not satisfy the characteristics mentioned in standard 3.3 (detailed earlier) that explain clearly what the test specifications should define.

Table 4 General instructions of Grade 12 CNE Life Science section

The Chemistry exam:
- is a mean to evaluate the levels of acquired competences in Chemistry.
- consists of competences from the three domains (applying knowledge, designing an experiment, and mastery-communication).
- should cover most content domains of the G12CC.
- should be based on pedagogic teaching practice that balance the three levels of knowledge (acquisition, transfer and production).
- is made up of three obligatory independent exercises formed of essay test items where each exercise has a title and may include one or more content domains.
- consists of exercises that connect scientific knowledge to every day life.
- is marked on a total of 20 points where the score of each of the three exercises is between 6 and 8 points.
- has a weight 4 in the LNE; thus, the total score attributed to Chemistry in the LNE is 80 points.
- is corrected based on a scoring rubric to insure consistency in correcting different copies.
- should be performed within two hours.

Adapted from NCERD (2000) and Lebanese Ministry of Education (July 2001)

As mentioned earlier, the first condition to develop parallel forms of an exam equivalent in content and cognitive demands is to construct them based on the same test specification includes a clear test blueprint (AERA et al., 1999; Newton, 2007). Since the grade 12 CNE were constructed based on very brief test specification and without test blueprint, it is extremely difficult for them to be parallel versions with same difficulty level. Therefore, the results obtained in this study were expected by the researchers and did not surprise them.

Indeed, all Lebanese national exams are designed based on insufficient test specifications and without test blueprints. Therefore, there is a large probability to find that the Lebanese national exams used as parallel form are not equivalent in content and in cognitive demands.

Recommendations of the study

This research is relevant for Lebanese education policy. To assure comparability as maintenance of examination demands across different grade 12 chemistry national examinations, it is vital to revise the test specification of the exams and develop a clear test blueprint that relates the content of the test items to the chemistry curriculum (for content comparability), and maps the cognitive complexity of the items (for cognitive comparability). Hence by identifying, delimiting and describing the cognitive demands of the chemistry national examinations, further research can be carried out to study the alignment between standards and content and cognitive demands of national examinations.

The emphasis in the new Lebanese science curriculum is on the knowledge of science, the investigative nature of science, and the interactions of science technology and society, but neglects science as a way of knowing (Boujaoude, 2002). This emphasis, however, is not reflected in the teaching practices prevalent at the present time in Lebanese classrooms. The reasons for this mismatch are numerous including the current examination system, the scarcity of supportive instructional materials, and most importantly the amount and quality of professional development activities available to teachers (Ayoubi & Boujaoude, 2006).

The more evidence gathered in this area, the easier the push for more reform in curricula. With this also comes better preparation for teachers in implementing tasks that require higher levels of cognitive demands.

A more comprehensive investigation of the cognitive demand of examinations can be connected to research on its correlation to student's achievement. Once that is made it will be easier to place better textbooks in the classroom that provide instruction based on content and cognitive complexity.

References

- American Psychological Association (APA), and National Council on Measurement in Education (NCME)] (1999). *Standards for educational and psychological testing*, Washington, DC: Author.
- Anderson, JR. (2005). Human symbol manipulation within an integrated cognitive architecture. *Cognitive Science*, 29, 313-342.
- AQA (Assessment and Qualifications Alliance) (2005). *A basic guide to standards setting*, Guildford: AQA.
- Ayoubi, Z. & Boujaoude, S. (2006). A profile of precollege chemistry teaching in Lebanon. *Eurasia Journal of Mathematics, Science and Technology Education*, 3 (2), 124-143.
- Baird, J. (2007). Alternative conceptions of comparability. In Newton, P., Baird, J., Goldstein, H., Patrick, H. and Tymms, P. (Eds.) *Techniques for monitoring the comparability of examination standards*, London: QCA.
- Boujaoude, S. (2002). Balance of scientific literacy themes in science curricula: The Case of Lebanon. *International Journal of Science Education*, 24, 139-156
- Bridge, P. D., Musial, J., Frank, R.,Roe, T.& Sawilosky, S. (2003) Measurement practices: Methods for developing content-valid student examinations. *Medical Teacher*, 25 (4), 414-421.
- Downing, S. (2006). Twelve steps for effective test development. In S. Downing and T. Haladyna, (ed.) *Handbook of test development*, London: LEA.
- Dunn, L., Morgan, C., O'Reilly, M., & Parry, S. (2004). *The student assessment handbook: New directions in traditional and online assessment*, London: Routledge Falmer.
- Hess, K. (2010). *Applying Webb's Depth-Of-Knowledge (DOK) in science*, retrieved http://www.nciea.org/publications/DOKscience_KH11.pdf

- Jones, P., Smith, R. and Talley, D. (2006) Developing test forms for small-scale achievement testing systems. In S. Dawning and T. Haladyna, (ed.) *Handbook of test development*, London: LEA
- Kellaghan, Thomas; Greaney, Vincent; Murray, T. Scott. (2009). *Using the Results of a National Assessment of Educational Achievement*. *World Bank*. Retrievedfrom:https://openknowledge.worldbank.org/handle/10986/2667
- Lebanese Ministry of Education (June 2001). Decree number 5697. *Nizam al-imtihanat alrasmiya* [Regulations of National Exams] Beirut
- Linn, R. (2006). The standards for educational and psychological testing: Guidance in test development. In S. Dawning and T. Haladyna, (ed.) *Handbook of test development*, London: LEA.
- Murphy, R.J.L. & Broadfoot, P. (1995). *Effective assessment and the improvement of education: A tribute to Desmond Nuttall*. London: Falmer Press.
- NCTM (National Council of Teachers of Mathematics) (1995). *Assessment standards for school mathematics*. Reston, VA: Author., 1995
- NCERD (National Center for Educational Research and Development) (2000). Evaluation *guide: Chemistry - Physics (Secondary cycle)*, Beirut: Author.
- Newton, P. (2007) Contextualizing the comparability of examination standards. In Newton, P., Baird, J., Goldstein, H., Patrick, H. and Tymms, P. (Eds.) *Techniques for monitoring the comparability of examination standards*, London: QCA.
- Popham, H. (1999). *Classroom assessment: what teachers need to know*, London: Allyn and Bacon.
- Robinson, C. (2007). Awarding examination grades: Current processes and their evolution. In Newton, P., Baird, J., Goldstein, H., Patrick, H. and Tymms, P. (Eds.) *Techniques for monitoring the comparability of examination standards*, London: QCA.
- SCAA (School Curriculum and Assessment Authority) (1996). *Standards in public examinations 1975 to 1995: A report on English, mathematics and chemistry examinations over time*. London: Author.
- Webb, N. L., Alt, M., Ely, R., Cormier, M. and Vesperman, B. (2005a) The

WEB alignment tool: Development, refinement, and Dissemination, In Council of Chief Stat Officers (Ed.) (2006) *Alignment assessment to guide the learning of all students: Six reports*. Washington, DC: Author.

- Webb, N. L., Alt, M., Ely, R., Cormier, M. and Vesperman, B. (2005b) *Web alignment tool (WAT): Training manual*, Wisconsin: Council of Chief Stat Officers and Wisconsin - Centre for Education Research
- Wolf, A. (2000). A comparative perspective on educational standards. In H. Goldstein & A. Health Eds.), *Educational standards* (pp. 9-30). Oxford: Oxford University Press for The British Academy.