Higher Order Thinking in Chemistry Curriculum and its Assessment

Peter J. Fensham & Alberto Bellocchi

Queensland University of Technology

Victoria Park Road, Brisbane, Queensland, Australia

Abstract: Higher-order thinking has featured persistently in the reform agenda for science education. The intended curriculum in various countries sets out aspirational statements for the levels of higher-order thinking to be attained by students. This study reports the extent to which chemistry examinations from four Australian states align and facilitate the intended higher-order thinking skills stipulated in curriculum documents. Through content analysis, the curriculum goals were identified for each state and compared to the nature of question items in the corresponding examinations. Categories of higher-order thinking were adapted from the OECD’s PISA Science test to analyze question items. There was considerable variation in the extent to which the examinations from the states supported the curriculum intent of developing and assessing higher-order thinking. Generally, examinations that used a marks-based system tended to emphasize lower-order thinking, with a greater distribution of marks allocated for lower-order thinking questions. Examinations associated with a criterion-referenced examination tended to award greater credit for higher-order thinking questions. The level of complexity of chemistry was another factor that limited the extent to which examination questions supported higher-order thinking. Implications from these findings are drawn for the authorities responsible for designing curriculum and assessment procedures and for teachers.

Keywords: chemistry assessment, higher-order thinking in chemistry, chemistry curriculum, PISA questions, assessing higher-order thinking
1. Introduction

The development of thinking skills, or more specifically higher order thinking (HOT) skills, in school science is a desirable educational goal that features regularly in educational reform agendas, science curriculum documents, and the science education literature (Gallagher, Hipkins & Zohar, 2012). However, there are concerns internationally that dominant assessment practices in science focus on low order thinking skills and that this in turn encourages teachers to focus on pedagogies that emphasize rote-learning (e.g., Osborne & Dillon, 2008). This suggests that the assessment instruments associated with chemistry curricula play significant roles in supporting or hindering the intent for HOT. This paper is concerned with the degree to which HOT is expected in high school chemistry curriculum documents and to what extent examinations support or discourage HOT.

2. Approaches to Higher Order Thinking

Systematic research interest in HOT originated in the contribution of Bloom (1956) who suggested a hierarchy of intellectual skills based on six verbs. The lower three were recall, comprehend and apply, and the upper three were analyze, synthesize, and evaluate. Bloom’s Taxonomy inspired one of us (Fensham, 1962) to begin his research in science education by reanalyzing the results of a set of university chemistry students he had taught. These students had undertaken an assessment test that required them to answer five questions (20 marks each) out of eight. Three of the questions were designed to require HOT based on Bloom’s Taxonomy. The students who chose to answer these HOT questions had a bipolar score distribution compared with the more normal distributions for the other five questions. This turned out to be due to the HOT questions being avoided by high achievers and being attempted by low achievers who failed to recognize their difficulty. In the reanalysis of the
overall student rankings, greater weighting was given to the scores on the HOT questions and this led to a considerable reordering of the ranked scores across the set of students.

Subsequent authors have developed different hierarchies from Bloom’s Taxonomy to indicate the levels of reasoning or understanding in student responses to assessment instruments involving open-ended questions. The Structure of Observed Learning Outcomes (SOLO) taxonomy of Biggs and Collis (1982) has five levels of increasing complexity from pre-structural to extended abstract, that depend on how the student’s response (for level 1) involves single or unrelated pieces of information or (for level 5) integrates numerous pieces of information and then applies the integrated information to new or untaught situations (level 5). The verbs in the presenting question can be indicative of the level of response expected. For instance, analyze, apply, argue, compare/contrast, criticize, explain causes, relate, and justify encourage level 3 and create, formulate, generate, hypothesis, reflect, and theorize encourage level 4. In addition to the use of verbs, the SOLO taxonomy also integrates the task complexity to the verb descriptors so that student responses that contain only single or unrelated pieces of information are deemed to be unistructural responses. Students’ responses that integrate numerous pieces of information and then apply the integrated information to new or untaught situations are deemed as abstract relational responses.

Other taxonomies based on cognitive levels have been developed as tools to assist in the alignment of curriculum objectives with assessment and teaching practices (Webb, 1997; 2007). In this alignment procedure Depth of Knowledge in both the objectives and the assessments is indicated by four levels of mental processing: Level 1 (recall), Level 2 (skill/concept), Level 3 (strategic thinking) and Level 4 (extended thinking). The kinds of activities that students could perform to demonstrate competence at level 4 included “(a) developing and proving conjectures, (b) designing and conducting experiments, (c) making connections between a finding and related concepts and phenomena, (d) combining and
synthesizing ideas into new concepts, and (e) critiquing experimental designs” (Webb, 2007, p. 13).

The analytical framework that informed our study is derived from the measures for scientific literacy used in the OECD’s Programme for International Student Achievement (PISA) project. These assessment measures were employed with 15-year-old students and had characteristics that reflect each of the aforementioned taxonomies (i.e., Bloom’s, SOLO and Webb’s). In this PISA Science project, a clear distinction is made between knowing a science concept, principle, or procedure, and the active application of that knowledge to unfamiliar situations. This distinction is perhaps also implied in curriculum statements which commonly use the two words, “Knowledge and Understanding,” as a learning objective. A set of test items in the PISA project was developed for each of three scientific competences: *explaining science phenomena*, *investigating science phenomena*, and *using scientific evidence*. The test items were intended to present different cognitive levels of application of science and technology in unfamiliar contexts. The items included a mixture of item modes including single multiple choice, complex multiple choice, and open constructed answers. In practice, the different degrees of item difficulty were determined *post hoc* by the percentages of students succeeding with an item. Six levels of difficulty were established with 2% of the very large student sample from the many participating countries achieving the highest level 6 items and 90+% achieving the lowest level 1 items (OECD, 2006; 2009). Examples of the prose descriptors used to differentiate between levels 1, 3, and 6 are shown in Table 1.

Clear differences can be seen for these three levels in terms of how the cognitive verbs (identify, explain, apply) interact with different amounts of information and the familiarity of their applications. This provides an operational definition that goes beyond the mere use of verbs as in Bloom’s Taxonomy. For example, at level 6 the specification of “complex life...
situations” and using different sources of information to justify decisions on an issue offer elaborations about the kinds of student responses that are coded at this level. The PISA levels thus also reflect aspects of the SOLO taxonomy in basing the classification of student responses on the number of links made between different pieces of information.

Although the PISA scale reflects aspects of the HOT represented in the Bloom and SOLO frameworks, no provision is made for questions that require multiple steps to arrive at a solution. In our study of the HOT demand in senior chemistry examinations, some quantitative problems occurred that involved multiple steps to arrive at an answer, whereas other questions required students to combine knowledge from different topics in chemistry (e.g., REDOX and stoichiometry) to arrive at an answer. These question types did not fit neatly into our PISA categories. Given the amount of chemical content and the algorithmic nature of the questions about it, we categorized them as lower order items because they could only be answered with recall of rehearsed knowledge.

3. The HOT Legacy in International Science Curricula and Assessment

HOT has featured consistently in reform agendas in science education internationally (Osborne & Dillon, 2008). Recommendations for improving science education across the European Union (EU) included the following:

... science courses that engage students in higher-order thinking which includes constructing arguments, asking questions, making comparisons, establishing causal relationships, identifying hidden assumptions, evaluating and interpreting data, formulating hypotheses and identifying and controlling variables. (Osborne & Dillon, 2008, p. 24)
The legacy of the early taxonomies remains evident in the use of the cognitive verbs or their derivatives (e.g., evaluating, formulating) that are found in statements of reform such as the one above. Osborne & Dillon (2008) also drew attention to dominant assessment practices across the EU that emphasized low-order thinking, such as recall of factual information, leading teachers into pedagogies that were focused on rote learning. Osborne & Collins (2001) earlier were able to claim that assessment remains part of the “forgotten landscape” of science education research, which is unfortunate because of the capacity of assessment to support or hinder a curriculum’s intent for HOT by determining the pedagogies employed in school science classrooms (Fensham, 2006; Liang & Yuan, 2008).

In China, a comparison of the curriculum guidelines for 12th grade Physics and the test content of the exit examinations revealed an overemphasis on low levels of knowledge recall and application (Liang & Yuan, 2008). The curriculum and examinations did not encourage or require creativity, critical thinking, or the ability to conduct scientific enquiry; skills commonly associated with HOT. The merit pay system in China potentially reinforces a focus on these low levels of cognitive demand in pedagogy and learning because it rewards teachers whose students achieve well in the examinations. These findings are reinforced by research that is more recent across three countries (Israel, New Zealand, & Northern Ireland) and which found that high-stakes summative assessment and accountability measures hindered the curriculum intent of developing students’ thinking skills (Gallagher, Hipkins & Zohar, 2012). Reports such as Liang and Yuan’s and Gallagher et al.’s (2012) studies are unsurprising given that an entire model of measurement-driven instruction was designed exactly to use assessment to modify teaching practices (Cizeck, 1993). If standardized instruments focus on lower order thinking, then teachers are most likely going to focus their pedagogy on these goals. On the contrary, it can be anticipated that if assessment is designed
to promote HOT then this should influence teaching practices due to the “teaching-to-the-test” effect in standardized high stakes assessments.

Fensham and Rennie (2013) drew on a number of international examples in which teacher-generated assessment instruments provided the scope for attaining various HOT aspects of science curriculum intent that included connecting science and technology, involving students in decision making, and developing knowledge about the nature of science. Zoller (1993) has also reported success in examination questions requiring higher order critical thinking skills (HOCS) through the use of appropriate teaching strategies in undergraduate chemistry classes. From these two sets of examples it is evident that assessment regimes have the capacity to support or hinder any reform agenda for HOT or curriculum intentions for HOT. Furthermore, they suggest that the inflexibility of externally designed assessment instruments is more likely to hinder HOT, whereas the flexibility that is possible with internal instruments can support HOT.

In the context of this study, internal assessment refers to assessment instruments developed by school teachers and external assessment refers to assessment instruments (typically examinations) that are developed by assessment and curriculum organizations and administered to all students within a state or territory. The instrument inflexibility in an external test does enable achievement scores to be compared – norm referenced measurement. The flexibility that internal instruments can have makes such numerical comparisons difficult, but the students’ performances can be assessed and compared against a commonly described set of criteria – standards-based measurement.

A degree of internal assessment is a key feature in our study, therefore it is useful to indicate that we mean that the chemistry teachers in a school take responsibility for the design and grading of their students’ work on a set of tasks that are carried out during the course of
study. In the high stakes assessments that we have analyzed, the nature of these tasks is externally prescribed in broad terms, as are the criteria for the award of the different grades. This type of teacher responsibility extends the assessment practices that science teachers regularly undertake in their classrooms to check the various learning tasks they set their students. A long-standing example of one of these tasks in science is practical investigative skills. These can only be assessed in a balanced way by the classroom teacher as argued by Black, Harrison, Lee, Marshall & Wiliam (2004). Yung (2006) has described how biology teachers in Hong Kong carried out this role in high stakes testing with reference to the issue of fairness and comparability. The intended learning outcomes that are now expected in contemporary science curricula extends more of this assessment responsibility to the classroom teacher (Fensham & Rennie, 2013), and our current study gives examples of how this extension is handled in chemistry courses in Australia.

4. Scope of the study: Curriculum Intent for HOT vs. Assessment Reality

Our study explores the extent to which the curriculum documents, assessment procedures, and the examination instruments for chemistry in four Australian states support or hinder the development of HOT in their students.

School education in Australia, as federal country, has been a constitutional responsibility of its six states and two territories. Chemistry is taught in these jurisdictions as a separate science only in the final two years of secondary schooling (i.e., 11th & 12th grades). In this study we focused on 12th grade Chemistry because students’ learning is subject to summative assessment procedures that have significant consequences for students’ future options about further study or employment. The authority for Education in each state and territory publishes a range of curriculum documents about Chemistry to guide and assist the schools, the chemistry teachers, and their students. These documents include Key Information
about Chemistry, the Course Outline, and details about Assessment and Reporting. The four state authorities included the Queensland Studies Authority (http://www.qsa.qld.edu.au/), the Board of Studies New South Wales (http://www.boardofstudies.nsw.edu.au/), the Victorian curriculum and Assessment Authority (http://www.vcaa.vic.edu.au/Pages/index.aspx), and the South Australian Certificate of Education board (http://www.sace.sa.edu.au/).

We chose four states that are known to have both similarities and differences in assessment practices that were likely to be relevant to the presence or absence of HOT. These states also span the spectrum of how external and internal assessments of chemistry learning are used to express what the students have learned. New South Wales (NSW) and Victoria (VIC) are known as states that make substantial use of external examination of their students’ learning. South Australia (SA) makes more use of internal assessment, while retaining an external examination in the 12th grade. Queensland (QLD) has not used external assessment to assess learning for many years. However, the internal Supervised Assessment (see Table 2) can, and often does, take the form of a school-based examination (QSA, 2004; 2007). The procedures for assessment of chemistry in Western Australia, Tasmania, Australian Capital Territory and Northern Territory lie with this spectrum.

Table 2 summarizes the roles of external and internal modes of assessment in the four states that were the focus of our study, and describes the types of internal and external assessment instruments, their weightings (%) in determining overall student achievement, the types of criteria to be addressed (e.g., knowledge, investigation skills), and how the grading of assessment responses is denoted by either letter grades (i.e., A, B, C, D or E) or numerically (marks) for each instrument. The degree of control that schools have in determining the characteristics of the internal assessment varies significantly across the four states. For
example, in NSW the weightings (as %) that internal assessment should make to each of the three criteria (*knowledge, investigation & communication*) is prescribed, but individual schools decide the tasks used to assess these criteria. A set of tasks is suggested in the curriculum document that include: (i) investigation using secondary sources; (ii) planning and performing a practical task; (iii) performing and reporting a firsthand investigation; and (iv) an examination.

We present our analyses for HOT in two sections. In the first section, we present the analysis of the four curriculum statements of intent for chemistry, the key information to schools, teachers and students about the chemistry course of study. In the second section, the analysis of the actual examinations and other assessment procedures are presented.

5. **Analysis of Curriculum Intent**

5.1 **The Key Information**

A content analysis of the key curriculum information from each of the states was conducted to identify the extent to which HOT was emphasized in the intended curriculum. Such key information reflects typically two dimensions of a curriculum: vision and structure (Wiles & Bondi, 1989). The vision represents a philosophy of education that drives the development of the structure of the curriculum, which is the way in which the content is organized. In the curricula that we investigated, the vision was found typically under sections called “Rationale,” and “Aims and Objectives.” A common aspect of the structure of the state curricula was that the science content was listed in three sections, knowledge and understanding of science, investigation skills, and communication or evaluation skills.

Terminology that reflected HOT was identified in all four curricula in both the vision statements and the detailed structure of the documents. Verbs such as *evaluate, assess* and
justify, reflecting aspects of HOT in Bloom’s Taxonomy, were found consistently in the rationales for all four statements of vision. Each state used other verbs like analyze, critically reflect, and problem solve that further imply intention for HOT. The detailed aims and objectives expressed a developmental sense for both the students’ learning and its association with the progression in the content demand of chemistry. For example, there were statements of intent about developing students’ capabilities as global citizens as well as furthering their understanding of chemistry concepts and skills such as problem solving.

5.2 Analysis of Assessment Instructions in the Curriculum

Specific intentions to develop HOT were expressed in the information and instructions about assessment provided in the curriculum documents. Each state identified science investigations as a vehicle for developing HOT skills – develop questions, critically evaluate, think critically and reflectively, construct hypotheses, analyze and evaluate, and solve problems. The skill of communicating chemical information and communicating to a range of audiences is also a common curriculum intention. Together, the Key Information and the Assessment Instructions provide evidence of strong interest in HOT in each of the state’s intentional statements for Chemistry.

The instructions for VIC and NSW set out mark bands for their internal assessment. In NSW six achievement bands are specified and a student who achieves marks between 90 and 100 on the test is designated band 6 level and a mark below 50 is band 1. Each band is associated with a standards descriptor. For example, the band 6 standards descriptor represents student responses that include:

- demonstrate[ion of] an extensive knowledge and understanding of the concepts of the chemistry course content including context, prescribed focus areas and domain
display [of] an outstanding ability to describe and explain chemistry concepts, including abstract ideas, clearly and accurately, and to apply the concepts to unfamiliar situations

The verbs demonstrate and display that are used in the band 6 descriptor do not reflect verbs associated with HOT, despite band 6 representing the highest level of achievement expected. This suggests that the internal assessment instruments are unlikely to support HOT if their design is informed by this standards descriptor.

SA and QLD refer to assessment in terms of levels of achievement, and QLD elaborates these in detail for five achievement levels, A-E\(^1\). SA sets out the characteristics of an A, B, C, D, and E grade level for each of its four course criteria using three or four descriptors (see Appendix A), and these characteristics are available to students. These standards are applied to both internal and external modes of assessment. In QLD, the curriculum authority publicly mandates the standards to be associated with the content criteria. For example, to achieve the A level in the knowledge and understanding of chemistry criterion, a student’s work must demonstrate the following standards:

- reproduction and interpretation of complex and challenging concepts, theories and principles
- comparison and explanation of complex concepts, processes and phenomena
- linking and application of algorithms, concepts, principles, theories and schema to find solutions in complex and challenging situations. (Queensland Studies Authority, 2007)

Terms such as linking, application, interpretation and explanation are indicative of HOT. The use of the qualitative adjectives, complex, and challenging, also opens the possibility of

---

\(^{1}\) In the five levels, A, B, C, D and E. Passing grades is a C and above (i.e., A & B), and D and E represent failing grades.
rewarding HOT responses. This use can also promote the design of HOT questions because, as Webb (2007) pointed out, assessment instruments should be aligned to the expected achievement standards. The standards descriptors from NSW, in contrast, would suggest that their assessment instruments, if well aligned, are likely to target lower order thinking skills rather than HOT.

5.3 Analysis of Examinations

We now turn to the analysis of the examinations used in the four states in order to determine the extent to which their question items reflect HOT levels. The examinations that were analyzed were the publicly available external examinations for 2010/2011 in VIC, NSW and SA, together with an examination example from QLD that fits within its Supervised Assessment category.

5.3.1 Analysis Method for HOT in Examinations

Our operational criteria for analyzing the HOT demand in the examination question items were based on the PISA Science descriptions of six levels of difficulty for three scientific competences: explaining science phenomena, investigating science issues, and using scientific evidence (cf. Table 1). Many of the question items in the examinations could be related to these PISA descriptions. The groups of items that did not fit these competency tasks were those that required students to carry out quantitative calculations (e.g., stoichiometry) or to work through a number of steps in an organic synthesis. By analogy with the PISA finding that the difficulty of the task increased with the number of discrete ideas requiring linkage, we added to the PISA levels a consideration for these quantitative items and multi-step items based on the number of steps involved. For example, the question extracted from the NSW examination was rated as a level 5/6 on our modified PISA scale because there are multiple steps involved at arriving at an answer.

[Figure 1 Here]
The two of us independently classified each item (i.e., examination question, N= 229) on the four examinations using the modified PISA Science classification. In 76% of cases our classifications were consistent and when the six PISA levels were compressed into three (i.e., level 1/2; level 3/4; level 5/6) the agreement was even higher. For the few items where our classifications differed, discussion of the depth of chemistry involved and what was demanded in a student response to the item in relation to the PISA levels led to agreement. Because our interest was in items requiring a HOT response, we decided to use only the level 5/6 classifications as indicative of HOT and levels 1-4 as indicated of lower-order thinking (LOT). Level 5/6 items were generally so rare that this approach to an agreed classification was highly effective.

6. Findings

6.1 Examinations

Each state devotes the majority of its examination question items (and marks) to aims and objectives about “knowledge and understanding.” In VIC the question items are almost all of this kind, with only 2 of the other 11 aims having any representation. NSW gives significant attention to “planning investigations” and to “applications of chemistry” leaving only 3 aims and objectives out of 14 not represented. SA does not include 2 of its 6 aims and objectives and QLD includes “evaluation and concluding” and “investigative processes.” The extent to which the internal assessment compensates to give more balance to the course aims and objectives is discussed in section 6.2.

The detailed findings of our analysis for the presence of HOT in the four written examinations are presented in Table 3.
The table presents the number of question items associated with our coding categories, for example, the VIC examination contains a total of 54 questions and 38 of these were coded as level 1/2 based on our scheme. In Table 3, “marks” refers to the numerical scores associated with question items. For example, a question is allocated “3 marks” by the examiner and a student is awarded 0-3 marks based on the quality of their response. Also shown in parentheses is the total number of marks allocated to all questions in the examination and how many marks were associated with questions pertaining to our coding scheme. In the VIC example, there were a total of 72 marks in the examination and 50 marks corresponded with questions in the level 1/2 category. Frequencies for our coding scheme were calculated by dividing the number of marks for each coding level by the total number of marks. For level 1/2 in the VIC examination the frequency is 50/72 = 69.4%.

As indicated from the first two columns of the table, the majority of marks in all four states were dedicated to questions in the 1-4 PISA categories. That is, the bulk of the examinations are devoted to items requiring LOT.

NSW had the highest number of level 5/6 items, but the marks allocated to these items (25 out of 200) were disproportionate to the HOT demand of the items. Furthermore, these HOT items amounted to only 9% of all items. A similar scenario occurs in the VIC examination with 2 marks out of 72 marks allocated to level 5/6 items (2% of total). Providing so little reward for these HOT items could serve to discourage students from attempting to answer them. In contrast, QLD has 4 out of 20 of its items at 5/6 level (i.e., 20%) and these can be rewarded with higher grades (i.e., A & B) than in the other states. That is, the level of achievement awarded to a student in QLD reflects the HOT demand in an item. In the other states, it is possible for students to achieve high overall grades or marks by answering correctly a large number of low order thinking questions.
The QLD examination was not strictly comparable with the other states on several grounds. Firstly, the grading system was based on criterion-referenced levels (A-E) and not in terms of numerical marks. Student responses to question items were graded according to marking criteria rubrics by using one of the five levels A-E. The level achieved depended on whether the response demonstrated HOT as mandated in the criteria matrix. In contrast, the other states allocated the number of marks associated with each question item (e.g., Question 1=5 marks) and graded student responses numerically (e.g., response = 3 out of 5 marks). Secondly, the QLD examination was obtained from a publicly available example that was designed for 11th grade students, whereas the exams for the other states and territories were 12th grade (i.e., final year) examinations. This meant that the level of chemical content knowledge in the QLD examination was not as high as that of the NSW and VIC examinations that were designed for the 12th grade. This difference in chemical challenge did not complicate the analysis, because the categories for analysis were focused on the HOT demand of the items, not the level of chemistry involved. We addressed the role of chemical complexity separately to determine the level of challenge in examination question items.

6.2 Chemical concepts

There is considerable diversity in the extent of detailed chemical concepts included in the four examinations, with the order from most concepts to fewest concepts being VIC>NSW>QLD>SA. As noted in section 6.1, the QLD examination was designed for 11th grade students and this could account for the lower level of detailed chemical concepts involved. Associated with the number of detailed concepts in the examination items is an increase in the content demand of the chemical systems about which students are asked to respond. An example of this is presented in the following question from the VIC examination shown in Figure 2.

[Figure 2 Here]
The question requires that students work backwards from the formula \( \text{P}_2\text{O}_5 \) and the information provided in the stem of the question to determine the percentage by mass of phosphorous in a fertilizer. Multiple steps are required that involve different types of calculations in order to arrive at a final answer. The complexity in such a question comes from the multiple steps that are required to arrive at a solution. Students have to remember the different kinds of calculation involved for each step, and then sequence them correctly to produce a correct response. This question scores low on our PISA-based scale because it essentially involves recall and direct application of remembered algorithms. Nevertheless, students may find such a question difficult because of the amount of discrete information to be recalled and the nature of the chemical systems involved. That is, the chemical species (e.g., \( \text{MgNH}_4\text{PO}_4.6\text{H}_2\text{O} \)) in the question do not have simple chemical formulae and require more algorithmic manipulation than would be involved with simpler chemical species.

All four examinations included at least one item that related to the shared curriculum aim for Investigation Skills as seen in Figure 3. The question item in Figure 3 was regarded as a HOT item that addressed investigation skills relating to the use of data and involving the justification of two different interpretations of the same data set.

[Figure 3 here]
The emphasis placed on the curriculum aim, Investigation Skills, varied considerably in terms of the marks associated and the HOT level. The rank order for these items (marks and HOT level) is QLD>NSW>SA>VIC. Only QLD used items at level 5/6 to explore HOT for investigative skills. Consideration needs also to be given to the weightings or relative value allocated to external examinations compared with the other internal school assessment (cf. Table 2). In SA, the external examination contributed towards 30% of the students’ final grades with the other 70% of grades coming from the internal school assessment. Therefore, it may be the case that in SA the internal assessment is focused to a greater extent on student responses that can manifest HOT whereas the external examination focuses on the lower levels of recall and direct application. Conversely, in VIC the external examination represents 66% of the students’ final grades. The external examination contains 98% of level 1-4 items providing very little scope for supporting curriculum intent for HOT. This means that the internal assessment offered the main opportunity for assessment to support HOT in the VIC system. However, the internal assessment has an impact of only 34% in determining students’ overall grades in Chemistry, thereby limiting the impact of HOT items (if they are in fact included in the internal assessment) on levels of achievement. In NSW where the external examination contributes to 50% of the final grades there is a similar, but less pronounced effect, against HOT. In contrast in QLD, where 100% of the assessment is internal, the syllabus guidelines specify that all three different types of assessment tasks (i.e., investigation reports, written assignments, and examinations) must contribute in the determination of overall grades thereby reducing the impact of examinations on students’ achievement.

An assessment structure like the one used in QLD is able to provide greater support for HOT in two ways. The first supports HOT by limiting the number of examinations that comprise the assessment package used to determine students’ overall achievement. In the QLD and SA systems the complete assessment package includes a variety of tasks creating a
balance between investigation reports, written assignments and examinations. In these systems, the students’ final results are not skewed towards the result on examinations because all assessment tasks may contribute equally to final student results. Examinations involve predominantly short response items so that the examination can be administered to large numbers of students over short periods of time. HOT items can require extended time for students to devise a solution and forms of assessment such as essays, investigations and reports, can offer these extended times frames. This also allows for more open or complex problems to be administered.

The second way in which HOT is supported is through mandated marking criteria and standards that apply to all assessment types. The standards descriptors for A-E levels of achievement consist of mandated statements like the ones used by SA (cf. Appendix A). In criterion referenced assessment systems, the level A descriptor is based on HOT and the order of thinking decreases in each subsequent level (i.e., B-E). By mandating standards that reflect different levels of HOT progressing to LOT, teachers need to design assessment instruments that allow students to demonstrate the higher levels of response. In this system, it is not possible for assessment instruments to focus on recall alone because this would ultimately disadvantage all students as they would have limited opportunities for achieving higher levels such as A and B.

The findings in Table 3 and our analysis of the variation in the chemical concepts included in the examinations in 2011 across the four states present a very different picture from the one reported by Matters and Masters in their study of Year 12 Chemistry content and achievement standards (Department of Education, Science & Training, 2007). They found an 80+% similarity in the concept coverage for chemistry across the Australian states and, as far as their limited access to the respective assessment standards allowed, that these also seemed to be quite similar (Department of Education, Science and Training, 2007). In
contrast, our focus on the actual HOT level of question items and level of credit given to items that focus on HOT suggests that there is a large disparity in the standards expected by the four states. Superficially, the documentation for curriculum intent and the assessment standards specified by each state may suggest similarities, but the assessment items suggest clear differences in the reward that is given to HOT and hence, how students’ HOT contributes to determining students’ overall achievement in Chemistry.

6.3 Internal Assessment

The varying but significant proportion that internal assessment (i.e., developed by the teachers) contributes to a student’s overall learning assessment in each of the states raises possibilities for rewarding HOT that are not so easily achieved in the external/examination type of assessment. This is most possible in QLD with its 100% internal assessment, followed by SA with 70%, NSW with 50%, and is least in VIC at 34%.

Extensive instructions are provided in each state to schools, teachers and students about the design and application of the internal assessments for which they are responsible. An explicit emphasis on HOT is evident in the instructions for SA and QLD (South Australian Certificate of Education, 2012; Queensland Studies Authority, 2007) that is not so evident in the cases for NSW (Board of Studies NSW, 2002); and VIC (Victorian Curriculum and Assessment Authority, 2005). The first two states also publish examples of student work that illustrate an investigative task in practice and the levels of performance in their grading system (e.g., http://www.sace.sa.edu.au/subjects/stage-1/sciences/chemistry; http://www.qsa.qld.edu.au/1952.html#assessment). In each state investigative work involving primary data collection, secondary data analysis and reporting of investigations is stressed for the tasks students are to carry out for their internal assessment.

6.3.1. SA and QLD
In both SA and QLD teachers are required to design opportunities for students to provide evidence of learning at the highest possible level of achievement, and to report a student’s learning achievement in terms of the designated criteria for the course, using five grade levels, A to E. In SA the criteria include “investigation”, “analysis and evaluation”, “application” (new and familiar contexts) and “knowledge and understanding” (use knowledge to understand and explain social or environmental issues) as shown in Appendix A. In QLD they are rather similar – “knowledge and conceptual understanding” (ability to acquire, to engage with, and to interpret....), “investigative processes” (ability to recognize methodologies, to judge worth of qualitative data, and to interpret and apply the outcomes of quantitative data), and “evaluating and concluding” (ability to synthesize their thoughts and others’, to determine interrelationships, to propose solutions and justify decisions, and to communicate their findings). These assessment criteria apply to each type of internal assessment (SA, 2 tasks; QLD, a minimum of 3 tasks).

6.3.2 NSW and VIC

Both these states again stress their interest in high levels of student achievement and refer to “breadth and depth of learning.” Teachers are provided with band descriptions of performance in terms of marks to assist their grading of the internally assessed work of students. These band descriptions are not available to the students, only a statement of the assessment tasks to be undertaken and their mark weighting. Illustrative examples of students’ work in the assessment tasks are likewise not publicly available. A student’s final internal grade is not a descriptive or A-E grade level, but a mark that reflects the assessment over the various internal tasks (NSW, 3-5 tasks; VIC, 3 tasks). This mark is then subject to statistical moderation against the school’s performance on the external examination that is itself weighted to lower order recall of chemical content in both states – NSW (53-64.5% depending on optional topic) and VIC (90+%).
In NSW the 3-5 tasks on which the internal assessment is based are required to include the fourteen learning objectives for chemistry, and these are grouped with sub-weightings that ensure that overall there is a weighting of 40% to “knowledge/understanding”, 30% to “investigative skills”, and 30% to “skills in communication/scientific thinking/problem solving.” Together with the external examination, the internal assessment places further emphasis on knowledge/understanding by requiring a 40% contribution to student grades. However, the internal assessment offers scope for assessing HOT through the “investigative skills”, and “skills in communication/scientific thinking/problem solving” objectives.

In VIC, where the external examination focuses almost entirely on the aim of student’s understanding of factual chemical knowledge, the internal assessment tasks are responsible for 6 of the subject’s other 11 aims that refer to “investigative skills and communication”, but with a combined weighting of at most 34%. Even so, the sharp differentiation in VIC between the roles played by the two forms of assessment still leaves four subject aims for chemistry unaddressed by either the external or internal assessments, and two of these have obvious potential for HOT. These two aims are to understand the ways chemical knowledge is organized, challenged, revised and extended, and to assess the quality of assumptions and the limitations of models, data and conclusions. The first of these aims supports HOT because it refers to the way in which chemical knowledge could be extended. This suggests that students should be asked to think beyond the received facts of chemistry to future chemical explanations of phenomena or to think about possible revisions of existing chemical explanations. The second of these aims also supports HOT because it requires the critical evaluation of models, data and conclusions.

7. Supporting Curriculum Intent for HOT Through Assessment: Lessons Learned

7.1 Curriculum Statements of HOT intentions
NSW and VIC both stress an interest in HOT in student achievement and refer to “breadth and depth of learning,” whereas SA and QLD refer to student learning at the “highest possible level of achievement.” It is clear from our study that simply providing such statements about the HOT intent in curriculum documents is not sufficient to ensure they will be heeded in the assessment of student learning. The positive intention for HOT can easily be lost if the assessment procedures do not specifically allow opportunities for students to exhibit and be rewarded for HOT in their responses.

In each state the course aim/objective about the acquisition of knowledge/understanding of chemistry is associated with a majority of the question items (and the marks in VIC, NSW and SA) on the external examination, but only in NSW and QLD is this operationalized to explore HOT. For example, the verb explain that is often associated with HOT in the literature, did not require a high level response in almost all the question items we analyzed. In most cases, questions that invited explanation could be answered by an associated recalled “fact.” Ogborn, Kress, Martins and McGillicudy (1996), have documented how commonly this low level of meaning for explain is used in science classrooms. In its more usual scientific sense, explain involves a different level of analysis of an original fact or finding, such as the three levels of chemical thinking (macroscopic, microscopic, and symbolic) promoted by Johnstone (2000) and Gabel (1999), but no examples of this occurred in the examinations.

7.2 Opportunity and Credit for HOT

In VIC the main opportunity for HOT was within the internal assessment tasks, and this is constrained by the low weight given to them and by the statistical moderation of the marks scored against the external examination with its minimal recognition of HOT items. Furthermore, if a student’s performance in the internal tasks is at a HOT level, the credit for
this is subsumed into the overall mark making it less visible in their overall achievement. In contrast, the use of HOT descriptors in the criterion referenced systems in QLD ensure high visibility of HOT in determining students’ overall achievement. This is because the mandated standards in the criteria matrix for the A level of achievement is based on HOT. Opportunity for students to exhibit HOT fares better in NSW than VIC, as it is present within both its external and internal assessments, but even so, a student’s performance on HOT items is lost when the marks are combined to give the total assessment score after moderation. The external examination in SA is limited in its HOT demand, but the internal assessment tasks may give opportunity for HOT. However, the standards descriptors associated with the Knowledge and Understanding marking criteria in SA were focused on LOT. In QLD all three internal assessment tasks provide opportunity for HOT and its criterion-referenced system credits this level of performance accordingly.

7.3 Coverage and complexity of Chemical Concepts

The variation in Table 3 of marking reward for HOT can be related to the extent of chemical coverage included in the questions and to the complexity of this chemistry. If the examination is designed to include all or most of the concepts that have been explicitly taught in a semester or year-long course, then there needs to be a large number of items, all or most of which carry only a small number of credit “marks.” Furthermore, a large number of items must be in a format that can be quickly responded to, such as simple multiple choice or direct short answer. The external examinations in VIC and NSW give priority to this type of coverage and to these types of item. These examinations provide discrimination across the student population in terms of students being able to recall more or less of the chemical concepts covered in the teaching.

There are alternatives that maintain high coverage in the examination such as making clear that a certain percentage of HOT items (and marks) will be used to cover each major
chemical topic. Open response items facilitate HOT but they do require more time. The “optional content” topics in the NSW examination could have been an opportunity for such higher crediting, but this was not used consistently across the options in the examinations we analyzed.

The chemical content knowledge presented in the examination questions, particularly in the examination in VIC, needs comment. More advanced chemical content may discriminate between the students’ achievements, even though low level questions (in terms of our HOT classification) are being asked about it. If this is so, to ask higher level questions about advanced chemical content would make answering doubly difficult because the student would need to recall details of the advanced chemical content and apply it to a more open problem.

NSW and QLD have shown that it is possible to ask higher-level questions about simpler chemical systems, although the marks reward in NSW is limited compared with the distinctive recognition that HOT questions gain in the grading system used in QLD. The more advanced chemical content in the examination items may also have the effect of exhausting some students’ capacity to recall the content, and hence will again differentiate them from others with greater ability to recall the large numbers of concepts, but not on their ability to apply HOT.

7.4 Applications of Chemistry

Given the familiarity in Australia of the contextual questions of the PISA Science tests, it is surprising that only NSW (consistent with its explicit objective about applications) included items that involve application of chemical knowledge to what seem to be unfamiliar contexts in the core of the external examination. We use the term context to refer to applications of chemistry to realistic scenarios. The PISA project has shown such contexts can extend the level of thinking required to answer a question (OECD, 2006). NSW (through its options) and
QLD do require substantial attention to be given to chemistry in context. In the case of the options in NSW the questions seemed to be associated with contexts that would be familiar from the teaching and study of the optional topics and hence were low level. In QLD the assessment of the context-based learning was by the other internal tasks (e.g., scientific reports) as well as the examination possibility.

8. Implications

8.1 Alignment issues

The findings from the study of the role of assessment in chemistry in the senior years of schooling raise a number of wider issues about how assessment practices can support or hinder the explicit intentions about HOT that are now regularly to be found in policy/curricular documents. The majority of these are issues for the educational authorities responsible for the curriculum of a subject and for its assessment procedures. Where these lie within the same authority there can be slippage between the curriculum’s formulation and the group brought together to design the assessment instrument(s). Where different authorities are involved the brief for the assessment authority needs to be very explicit about how the HOT intentions are understood.

There are also implications for teachers of science subjects, but they are contingent on the specification’s from the curriculum and assessment authorities for the external assessment and especially for the rules stipulated any internal assessment.

8.2 Curriculum and assessment authorities

The implications for curriculum authorities responsible for Science can be posed as three questions – What sort of scientific literacy do they see their curriculum developing in young people and to what extent does it involve HOT? How serious are they about HOT levels of
learning becoming teaching/learning aims and intentions for Science and its associated subjects? and in conjunction with the assessment authority, How can it be ensured that the assessment and reporting procedures for the students’ achievements of these aims/intentions are supported?

8.3 Scientific literacy?

Roberts (2007) posed two visions of scientific literacy: Vision I refers to a science education aimed at preparing students to become future scientists; Vision II focuses on preparing scientifically literate citizens. Until full secondary education became the goal and norm for all young people, Vision I was the only role the sciences in the senior years had to play. For more than thirty years, however, Vision II has also been acknowledged as an additional important role for school science, and this has continued to tax authorities to design a curriculum and assessment procedures that will serve both roles.

There is agreement that science content knowledge and skills of investigation are seen as important for both visions. The range of this content and the HOT levels of its understanding are, nevertheless, often in contention, and this applies also to what students should know about scientific investigation. For example, each of the four chemistry curricula in our study differ in terms of the amount of chemical content included and the HOT level of its required level. Content knowledge is central in the three external assessment tests, where their use is for a norm referenced purpose, but its HOT level is only credited in the case of the internally set and criterion referenced test in Queensland. The norm referenced purpose of a test does not have to, in itself, preclude HOT recognition. Indeed, until the 1960s curriculum reforms in assessment in these senior years involved a mixture of quantitative and extended prose questions that led to a measure of different levels of knowledge preparation. Since the 1960s there has been a rise of psychometric testing of performance dominated by curriculum
coverage and multiple choice items which has meant that the opportunity to encourage HOT learning has been reduced especially in norm referenced situations. The complexity and detailed coverage of chemical content included in the external examination in VIC and how this knowledge is credited reflects a predominantly Vision I perspective without HOT support.

Investigative knowledge is recognized in the four curricula, and to the extent that it is encouraged and the contexts in which it is applied means that it can serve both visions of literacy. It is, however, now increasingly being argued that Vision II, because for many students senior school science will be their last study opportunity, needs a deeper and wider understanding of how scientific knowledge is established and developed. Responses to this need are reflected in the attention now being given in the literature to scientific argumentation (e.g. Osborne, Erduran & Simon, 2003; Erduran & Jiménez-Aleixandre, 2007) and to the use of evidence in science (Bybee, McCrae, & Laurie, 2009). Stengers (2011) described the Vision II goal as a “connoisseurship” of science that certainly encourage a more discursive and sophisticated HOT understanding of science.

8.4 Aims/intentions

We have found aims/intentions for chemistry in these four curricula that are by no means evenly supported by the rewards associated with their assessments. Furthermore, when some intention is targeted by an assessment task its successful learning is lost because it contributes only a small number of marks to an overall score that is dominated by other aims/intentions.

It is now time for the authorities across Australia and elsewhere to consider expanding their present dedication of just some of their teaching/learning aims to several other aims/intentions that are not well served by the format of written examinations. Obvious ones from the curricula we investigated are:
• Evaluates how major advances in scientific understanding and technology have changed the direction or nature of scientific thinking (H1, NSW)

• Analyses the ways in which models, theories and laws in chemistry have been tested and validated (H2, NSW)

• Understand the role of experimental evidence in developing and generating new ideas and new knowledge in chemistry (2, VIC)

Decision making about socio-scientific issues has been listed for some years as an aim/intention in many countries’ science curricula and it is referred to in various wordings in each state’s curriculum statement. This aim involves HOT skills like weighing evidence and deciding between conflicting outcomes that have usually escaped the attention of most teachers because they have not been required for assessment and are not amenable to the dominant written examinations. Various ways of assessing this type of learning are now available if authorities are prepared to endorse the appropriate learning tasks and allow teachers to provide their assessment of them as a component of the student’s overall report.

8.5 Designing assessment

To avoid the token encouragement of a wide set of aims/intentions but the domination, in practice, of just some, curriculum authorities need to ensure that there are dedicated tasks and credit that reflect the aims/intentions they value. As we noted, in the QLD system the HOT intent in the curriculum is supported through the mandated criteria that teachers are required to apply to all assessment tasks.

McTighe & Thomas (2003) outlined a backward-design approach as a more general and useful starting point for curriculum authorities to combine aims/intentions and assessment. In this approach the desired aims/intentions are first identified, and then for each, appropriate and authentic assessment tasks and modes are selected that most relate to
intended learning of knowledge and skills. External assessment will emerge as appropriate for certain aims, but not for others. Internal assessment, with its greater flexibility in assessment tasks and modes of response, will be seen to support a range of other aims/intentions that are not amenable to an external written test. In the case of HOT skills, this approach would commonly lead to criterion-referenced assessment against specified common standards, and it would become an essential aspect of the curriculum’s design through that of the pedagogies of the classroom. In our study, QLD and SA provide examples of this.

8.6 Reporting of achievement

If authorities do wish to value a set of aims/intentions, it follows that their achievement by students should not be lost in a single composite score. A profile of achievement provides a much more valid and informative report of learnings that are not obviously commensurate. Such profiles are much more helpful for the provision of advice regarding further study and/or employment. Employers now often supplement the enigmatic single score on an applicant’s report with their own measures of applicants’ responses to interview. Different faculties in universities are increasingly making use of students’ profiles, rather than a single composite score of final year achievement in several subjects.

For example, since 1990 university selection in Queensland has been based on Three Parts:

- a band level (Overall Position) derived from the school-based profiles of year 11/12 subject achievement,
- a measure (Field Position) of a student’s skills in specific subject areas, that are based on four dimensions (extended prose writing, short answer writing, basic numeracy, and complex quantitative problem solving) that vary in importance across school subjects and, in turn, in their interest to university faculties.
- the student’s results in a Core Skills test.
Running head: HIGHER ORDER THINKING IN CHEMISTRY

The grading of the Overall and Field Positions immediately brings in professional implications for the subject teachers. When these assessments relate to high stakes accreditation like university entrance, they must be more than just a teacher-produced grading. They must be seen to be fair and comparable across all schools as Yung (2006) has discussed for the case of biology in Hong Kong. A well established approach (in the cases of Art, Technology and Music) to such fair and comparable assessment involves small rotating teams of teachers visiting several schools to provide comment on that assessments their teachers have suggested. It has proved very possible for teachers to present and explain to their peers the basis of their assessments, and to adjust them in the light of their peers moderating advice.

Wilson and Sloane (2002) define moderation as the process in which teachers discuss student’s work and the scores they have been given, to ensure that these scores are interpreted in the same way across the moderating teams. Such moderation for quality control in assessment has been well developed in Queensland and Maxwell (2010, p. 457) described it as “a process for establishing confidence in the quality of procedures and judgments”. Peer-peer moderation may seem an expensive process but when its pre- eminent benefit as professional development is included in the equation, the cost is very reasonable.

8.7 Teachers’ pedagogy

Webb (1997, 2007) extended the alignment between a curriculum’s objectives and its assessment to include pedagogy by referring to the need for their alignment with the teaching practices employed, in our case, by chemistry teachers. It is well documented that assessment is a major driver of the teaching and learning in science classrooms (e.g., Liang & Yuan, 2008). Teachers take personal initiative and responsibility about the pedagogy of their classrooms, and the teaching strategies they employ. If there were little reward for HOT in
their students’ assessment, it would be foolish for teachers to use pedagogies that develop HOT skills, at the expense of other pedagogies that relate more directly to what is being emphasized in high-stakes assessment.

On the other hand, if the curriculum and assessment for chemistry has been changed along the lines just discussed, teachers would need to become familiar with those pedagogies that research has shown to encourage various HOT learnings (Kolstø, 2000; Osborne, Erduran & Simon, 2003; Radcliffe, 1997; Sadler & Zeidler, 2008).

Zoller (1993) and Mitchell (2009) have both reported lists of pedagogical practices that have been found to promote HOT in chemistry students. In Australia, there is good evidence that teachers of chemistry at the senior level could fairly easily make this transition in their teaching. Many are likely to be familiar with a number of pedagogies for HOT learning that have been developed for use in the earlier years of schooling, when the teacher or teachers in a school are solely responsible for the assessment of learning (Baird & Northfield, 1992). In QLD where assessment responsibility has for many years rested within the school, chemistry teachers have learnt to develop tasks and examination items that can be answered at several levels of correctness and they, accordingly, use pedagogies (including sharing the meta-meaning of these levels with their students) that explicitly aim at developing an appreciation in their students of what higher order learning can be (Queensland Department of Education and the Arts, 2004).

9. References


Board of Studies NSW (2002). *Chemistry*. Sydney, Australia: Board of Studies NSW.


Running head: HIGHER ORDER THINKING IN CHEMISTRY


Running head: HIGHER ORDER THINKING IN CHEMISTRY


