Intelligent Performance Assessment of Students’ Laboratory Work in a Virtual Electronic Laboratory Environment

Ifeyninwa E. Achumba, Member, IEEE, Djamel Azzi, Victor L. Dunn, and Gloria A. Chukwudebe

Abstract—Laboratory work is critical in undergraduate engineering courses. It is used to integrate theory and practice. This demands that laboratory activities are synchronized with lectures to maximize their derivable learning outcomes, which are measurable through assessment. The typical high costs of the traditional engineering laboratory, which often militate against the synchronization of laboratory activities and lectures, have catalyzed the increased adoption of virtual laboratories in engineering laboratory education. The principles of assessment in the virtual learning environment are essentially the same as in the traditional learning environment, with the same requirements for fairness, reliability, and validity. This motivated the incorporation, in a Virtual Electronic Laboratory (VEL) environment, of a Bayesian network-based tool for the performance assessment of students’ laboratory work in the environment. This paper details a description of the assessment tool, its verification, evaluation (as an assessment tool within the VEL environment), and application processes.

Index Terms—Bayesian networks, laboratory work, performance assessment, sensitivity analysis, validation and reliability, virtual laboratory

1 INTRODUCTION

A critical role is accorded to practical activities in the undergraduate engineering (UE) course. Practical activities in the UE course have been classified into three types: Controlled assignments, experimental investigation, and projects [1]. Controlled assignments (also referred to as teaching laboratories) have been described as relatively short laboratory (lab) activities, devised and used by the instructor to enhance students’ understanding of taught concepts, to integrate theory and practice. In this context, controlled assignments will be interchangeably referred to as lab work or activities. The objectives of lab work have been highlighted by the authors of [1], [2], [3]. These objectives require that lab activities are synchronized with lectures to maximize their derivable learning outcomes, measurable through assessment. The typical high costs of the traditional engineering lab, which often militate against the synchronization of lab activities and lectures, have catalyzed the increased adoption of virtual labs in engineering education.

Assessment in the virtual lab environment is as critical as in the traditional lab setting. The need to integrate assessment tools into virtual lab environments is therefore imperative, in view of their increased adoption, especially in engineering lab education. The principles of assessment in the virtual learning environment (VLE) are essentially the same as in the traditional learning environment (TLE), with the same requirements for fairness, reliability, and validity [4]. Often, the tendency is to extend the traditional lab work assessment practices to the virtual lab environment. The traditional lab work assessment practices seem inherently deficient and subjective, with possibilities of bias and unfairness. This motivated the construction of a Bayesian network (BN)-based assessment tool for integration into a Virtual Electronic Laboratory (VEL) environment, for the performance assessment of students’ lab work in the VEL environment. Descriptions of the assessment tool, its verification, and evaluation within the context of the VEL are presented in this paper.

First, related work is presented in Section 2 and the VEL is highlighted in Section 3. Next, a brief overview of BN is given in Section 4. A description of the BN-based assessment model, in terms of its structure realization and parameterization, is given in Section 5. The evidential data extraction and evidence variable quantization and instantiation processes are described in Section 6, and the verification and evaluation of the model is detailed in Section 7. Section 8 outlines the model application processes, while Section 9, which gives the discussion and conclusion, highlights the contributions of the work presented in the paper.

1.1 Traditional Laboratory Assessment Practices

Engineering faculties often require students’ to keep records of their lab activities using journals and/or logbooks, and in some cases to produce written reports. Assessment of students’ lab work is often based on...
marking/grading these written evidence and/or students’ responses to questions relating to the lab activities. It is reported by Wilkins [5] that students’ lab journals, in their faculty, are inspected but not assessed. Assessment for each lab activity is based on marking a set of questions constituting part of the lab instruction sheet. McPhun [6] reported a similar assessment practice where, though students’ keep journals of all lab activities, assessment is based on marking a set of questions related to the performance and results of the lab work. SEEDCU [7] used the logbook as the primary mechanism for the assessment of lab work. A random selection of a set of lab activities (which may be different for each student) are marked. A similar practice reported by Pramanik and Ring [8] indicates that first a student’s logbook is marked in consultation with the student who is then given a chance to revise the logbook based on any adverse feedback received. The revised logbook is marked. The final mark for the lab work constitutes of the mean of the marks assigned to the original logbook, the revised log book, and two written reports. Dearden [9] reported an assessment scheme that is based on logbook and the physical observation of students as they undertake lab activities. The assessment by physical observation is based on a set of performance factors. Each factor has its own set of assessment criteria to be used in awarding marks. For example, “Intelligent approach” factor has the following four marking criteria: original, full of ideas, questioning each step, and not contributing at all (no ideas) [1]. The marks from the logbook and the physical observation contribute equally to the final mark. Jordana and Sánchez [10] reported an assessment scheme in which the final laboratory activity grade is made up of marks from written reports, portfolios, and individual lab exams.

These traditional lab work assessment practices seem inherently deficient and subjective, with possibilities of bias and unfairness. For example, the assessment scheme of [9] puts emphasis on attitude, qualities, and other factors that neither reflect the objectives of lab work nor allow them to occur [1].

1.2 Problems of the Traditional Laboratory Assessment Practices

Assessment practices based on written evidence have been criticized as failing to address espoused aims [2]. The criticism is not necessarily indicative of the questionable value of the students’ written reports, rather, it emphasises the need for the assessment of students’ lab work from a holistic approach. “The requirements of formal lab reports, their structure, and the assessment methodology are often ambiguous” [11]. Moreover, assessment of written lab reports is challenging and demanding, especially for large class sizes. It is difficult to sustain this approach because of increasing student numbers. Large class sizes pose the additional problems of fair and consistent assessment. Also, students often “doctor” lab reports [2].

Furthermore, assessment of practical work based on written evidence focuses on assessing only the end product, neglecting the work process, and the specific practical abilities/skills (competencies), which are part of the learning outcomes of engineering education [12]. This is not in line with the principles of the prevailing educational paradigm that demands that students’ demonstrate their abilities/skills for lab work, and their knowledge and understanding of concepts addressed by lab activities. It is, therefore, imperative that students’ lab work assessment practices should change, not only in accordance with the prevailing educational theories, but also in response to statements of criteria for engineering education by Quality Assurance Agency (QAA) [13] and Accreditation Board for Engineering and Technology (ABET) [14]. This need for change toward a holistically based assessment of students’ lab work can possibly be met through performance-based assessment.

1.3 Performance-Based Assessment

“Performance (-based) assessment calls upon students to demonstrate specific skills and competencies and to apply the skills and knowledge they have mastered” [15]. It has been demonstrated as an effective way of evaluating students’ lab skills [16], and highlights what a student knows and can actually do [15]. Work at performance-based assessment, in the traditional lab, used the physical observation technique based on a list of performance criteria [9], [17]. Also, the principles of Objective Structured Clinical Examination (OSCE), a widely adopted physical observation-based performance assessment tool in healthcare education, has been proposed for the performance assessment of engineering students’ lab work by [18]. The authors proposed Objective Structured Technical Examination (OSTE) as an engineering version of the OSCE. These assessment schemes are space, time, and assessor intensive, logistically demanding, and difficult to apply to large class sizes. Bahri and Trevelyan [19] proposed the use of an online survey instrument to assess practical-intelligences (knowledge and skills) gained from engineering laboratory activities. The instrument consists of a set of domain-related tasks. For each task, there is a set of 10 to 20 response items, each describing a possible solution approach to the task. Students are required to rate the appropriateness of each response item on a 7-point Likert scale and their responses are scored. This assessment scheme is not performance-based and the authors acknowledged that constructing the assessment instrument is challenging.

The work presented in this paper is performance based. There are two aspects to the work: A BN-based model is presented as an assessment tool for the performance assessment of students’ lab work in a VEL environment; and the virtual observation assessment methodology is proposed as an alternative to the physical observation technique, for VLEs. Virtual behavioral observation is the collection and recording of behavioral data while subjects are engaged in activities in an interactive virtual environment [20]. The model harnesses the strengths of BNs to make inferences about students’ competencies and performances, from their observed behavior. Virtual behavioral observation is a common tool of user modeling [20], and is also used in student modeling. Virtual observation has also been used in other areas of study and applications including examination of moon phases [21], clinical studies [22], and teachers’ professional development experience [23]. Arrington [21] found virtual observations more effective than direct physical observations and produced better data.
2 Related Work

Intelligent assessment often entails the use of artificial intelligence (AI) in the implementation of an assessment model, through such techniques as fuzzy logic, neural networks, BNs, and so on. A BN-based model for the prediction and assessment of students’ current state with respect to continuing or dropping out of a study, in an open/distance informatics course, was presented by Xenos [24]. The prediction is made based on students’ actions in the environment such as use of computers, web-based material, textbook, internet, family status, and hours of work. The design of a BN-based model for the prediction of students’ performance in a test and the application of a BN-based model in educational computerized adaptive testing to assess students’ performance were reported by Corbett et al. [25] and Mislevy et al. [26], respectively. Collins et al. [27] applied BN in adaptive testing of multiple student latent traits in a single item-based test, using granularity hierarchies. Bayesian inference is used to propagate knowledge over the hierarchies. Each trait is either mastered or not mastered. Zhang et al. [28] presented a BN- and rule-based model for the assessment of students’ learning, using students’ knowledge map and analysis of their responses to tests items. The model was designed for the assessment of students’ software usage abilities and skills in a computer course. Liu et al. [29] reported a BN-based student model for assessing students’ performance in a VLE, from their web portfolios. BNs have also been applied in the assessment of learners’ characteristics [30], inference of learners’ goals [31], and inference of individual student preferences and learning styles (LSs) [32]. Feng et al. [33] stated that “Bayesian networks have also been used to investigate the results of skill hierarchies using real-world data in intelligent tutoring systems.” The work by Millán and J. Pérez-De-La-Cruz [34] and Millan et al. [35] were focused on diagnosis in student modeling based on the use of BNs and computer adaptive tests.

These assessment tools, though BN based, are designed for the Intelligent Tutoring/Adaptive System (ITS/ATS) environment, the common and initial area of application of intelligent assessment models. OnLine Assessment of Expertise (OLAE) [33], a BN-based assessment model, does not monitor students’ behavior nor does it assess students’ abilities for lab work [36], [37]. Literature search has highlighted previous work on assessment in VLEs and ITS/ATS but has not identified performance-based assessment in the context of virtual laboratories generally and VEL specifically. The assessment model by Noguez et al. [38] designed for a VLE consisting of an ITS and a robotics virtual laboratory was used in the ITS component of the VLE to estimate students’ knowledge of course themes. Also, the Universal Virtual Laboratory (UVL) by Duarte et al. [39], specifically focused on people with motor disability, does not integrate any performance-based assessment model for evaluating students’ lab work.

3 The Context: The VEL

The VEL presents undergraduate electronic engineering students with a workbench (an interactive Graphical User Interface (GUI), shown in Fig. 1), on which to practically construct and simulate electronic circuits. It is a tool students can use to undertake curriculum-based lab activities in a realistic manner. The components and devices required to build a circuit are laid out on the workbench, in containers. Components can be selected, picked up, and dropped or connected on a breadboard. A component connected on the breadboard can be removed and connections can be made using wires of different colours. Both power supply and function generator are provided on the workbench. The schematic of a breadboarded circuit and the simulation commands are captured, transformed into a netlist, and passed to the simulator. Simulation results are returned to the GUI and displayed on the results panel.

Students’ behavioral (interaction) events (mouse click streams and key presses), in the VEL environment, are tracked, captured, and logged. Events logged in the VEL include component selection, pickup and placement on the breadboard, dropping of picked up components, removal of a component connected on the breadboard, and search for a component in a component container. The behavior log is analyzed and performance data extracted for input into the assessment model. More details of the VEL are given in [40].

4 Bayesian Networks

A BN consists of a directed graphical structure (the nodes and the directed edges connecting the nodes) and the parameters (probability tables) for each node in the network. Each node represents a random variable or hypothesis and each directed edge (link) represents a relationship between the two linked nodes, thereby creating a parent—child relationship. Each node in a discrete BN has a finite set of possible values (states) and can only take one of its possible values at any one time. All child nodes (nodes with parents) have Conditional Probability Tables (CPTs) that model the combined impact of its parents, while each leaf node (nodes without parents) have Prior Probability Tables (PPTs) describing the prior knowledge about them. Essentially, a BN represents a probability distribution over the set of variables, \( U = (A_1, A_2, A_3, A_4, \ldots, A_n) \), in the network. The joint probability distribution, \( P(U) \), of the BN is \( P(U) = P(A_1, A_2, \ldots, A_n) \). This can be factorized, using the product rule, to yield: \( P(U) = \Pi P(A_i|Pa(A_i)) \).
where \( pa(A_i) \) is the parent set of \( A_i \). The product rule is also an application of the conditional independence rule. Inference is the algorithmic estimation of the probability of a target variable, \( X \), given evidence at node(s) \( A_i \).

BN was preferred for this application, over other AI techniques, because it has a sound mathematical basis, enables reasoning under uncertainty, and facilitates the update of beliefs, given previous beliefs and new observations. In addition, it affords opportunities for the visual representation of a model.

5 The Assessment Model

5.1 Framework for the Model Structure

The model is referred to as LAP (Laboratory Performance) model. The LAP model, an offline performance assessment tool, is built from domain expert knowledge. The model is constrained to the broad framework of learning as consisting of three main variables: Knowledge—acquiring and storing facts, rules, and principles about a concept; Understanding (meaning making)—understanding of expressed or implied interconnections and relationships between facts, rules and principles about concepts, in relation to specific tasks or problem situations; and Abilities/Skills—mastering procedures and techniques required to apply knowledge and understanding in specific task or problem situations, in a domain context. This framework, shown graphically in Fig. 2, represents the integrated assertions and statements of [13], [41], [42], [43].

Schwartz and Robbins [43] stated that experimental tasks result in observable behavior which reflects associated internal traits, based on which inferences can be made about the traits which cannot be observed or measured directly. Dember and Jenkins [42] asserted that internal traits, including abilities/skills, may combine to produce an instance of behavior, while [41] stated that abilities exist from their effects in terms of performance and are inferable from observable behavior. Gagne and Fleishman [41] also asserted that the acquisition of knowledge, understanding, and abilities/skills results in improved performance which reflects learning. This assertion is supported by the definitions of assessment by [44] as a generic term for a set of processes that measure the outcomes of learning, in terms of knowledge acquired, understanding developed, and abilities/skills gained.

Often, there is the tendency to differentiate between abilities and skills as internal general and specific potentials, respectively. In this context, we use abilities and skills jointly to mean context specific potentials and shall often be written as abilities/skills. Based on the framework of Fig. 2 and the assessment scheme of [45], the LAP model takes a behavior-observation-trait-inference perspective.

Lab instruction manuals and task assignment sheets were reviewed, students undertaking lab activities in the TLE were physically observed, and the help of domain experts elicited, to identify and factorize the cognitive abilities/skills necessary to successfully undertake lab activities at the undergraduate level of electronic engineering. Key lab activity tasks were identified and represented with the action verbs: Design, analyze, construct, modify, measure, plot, observe. In this context, the verbs refer to the abilities to: Design—devise the schematic of a circuit to meet a specified need(s); construct—connect electronic components together to create a circuit whose collective behavior meets a given specification(s); analyze—practically examine a given circuit to detect behavior patterns and draw conclusions; modify—alter or adapt a circuit to meet a new purpose, application, or specification; measure—determine the size or amount of, express as a number, quantity; plot—generate vector graphics using a set of experimental data; and observe—look at experimental result with a view to its interpretation and data extraction. These verbs are referred to, in this context, as the High-Level Abilities/ Skills (HLAS). The set of abilities are all cognitive abilities. Psychomotor and interpersonal abilities/skills are not considered relevant in this context.

A HLAS can be factorized to a set of Basic Ability Components (BACs) in accordance with the statement by Gagne and Fleishman [41] that performance activities can be described in terms of BACs, most of which can be considered as attributes of behavior. Table 1 highlights a HLAS and its associated BACs. A BAC can further be factored into lower level BACs (LLBACs). The HLAS, BACs, and LLBACs constituted network nodes and the edges (links interconnecting the nodes) were created as follows: Each LLBAC was connected by a directed edge to the BAC from which it was derived, and each BAC was connected by a directed edge to the HLAS from which it was derived, forming parent—child relations. Domain experts then rated the importance/relevance of the relationship of each parent node to its child node (represented by the directed edge from the parent to the child), on a 0 to 10 scale. A rating of zero meant zero importance/relevance, which implies the expert does not believe that a relationship exists between the two nodes. If more than one expert gave an edge a zero rating, that edge was deleted from the

![Fig. 2. The framework for the LAP model.](image-url)
network. In this way, the network structure was constructed, with Fig. 2 as the basic building block.

Other cognitive abilities considered to have contributory effect on laboratory activity performance are Spatial and Receptive Communication (SRC) abilities. Spatial ability is one of the most important abilities a student should possess to be successful in the engineering profession [46]. Receptive communication is the ability to understand instructions that is key to carrying out a task.

5.2 Model Overview

Fig. 3 represents a graphical view of the LAP model showing the network nodes (variables) and the directed edges connecting them. Often, semantics are used to give edges and their directions particular meanings, within the context of the domain, reflecting the relationships they model. The most common semantic is causality, which implies an asymmetric cause-effect relationship. Causality is not, however, a requirement for BNs [47]. The edges in a BN can represent other relationships, such as containment, ownership, part-of, prerequisite-of [34], or any relationship that has meaning within the context of the domain. For example, [27] defined the edges in their BN model as representing the relationship of “skill → subskill,” where the parent node represents a skill and the child node represents a subskill of the parent skill. In this contest, the edges model “subproficiency → proficiency” relationship. Fig. 3 is, essentially, a graphical description of how the nodes interact probabilistically (that is, it defines the probabilistic relations among the variables (nodes)). The numbers, names, and categories of the nodes in Fig. 3 are listed in Appendix A (Table A), which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TLT.2013.1. Each node has been named to reflect, as much as possible, the proficiency that it represents.

The categorization (classification) of the nodes is highlighted graphically in Fig. 4, which gives a more compact representation of Fig. 3. Each smooth edged rectangle represents a class with a list of the numbers of the nodes that belong to that class. The figure also shows how the classes are interconnected. Appendix A (Table A), available in the online supplemental material, gives a description of the classes. The HLAS that are not amenable to factorization are those that are directly observable from behavior in which case they also serve as performance indicators (because they contribute to the robust view of a student’s abilities/skills) and indices (because they are directly observable from behavior and do not need to be inferred from behavior) namely, nodes 17, 18, and 22. The memory/feedback node (node 35) was introduced into the model to facilitate incorporating previous performance estimates into...
current belief estimates, if desired. The Learning Style (LS) concept was considered an influential factor on lab performance because LSs are said to serve as relatively stable indicators of how learners perceive, interact with, and respond to a learning environment [48]. Each student’s LS was evaluated using the Index of Learning Styles (ILS) instrument by Felder and Silverman [49], and stored as part of historical data.

5.3 Knowledge Elicitation and Model Quantification

Quantification is the process of parameterizing the CPTs and PPTs of the nodes in a BN model. The different parameterization techniques, based on the source of knowledge required for the derivation of the parameters, include domain expert knowledge [37], historical data [50], empirical data [31], and a combination of expert knowledge and data [26], [51]. It may be expensive to conduct empirical studies for every student lab performance related ability/skill, and lack of existing historical data, cost, and time constraints make the domain expert knowledge technique most attractive. The work on LAP model was done with the committed participation of three domain experts (down from an initial nine) consisting of a Principal Lecturer, a Senior Lecturer, and a Reader, from the Electrical and Computer/Electronic Engineering (ECE/EEE) Departments of two different universities. Each expert had over 10 years of ECE/EEE student classroom/laboratory teaching experience. Their judgements were elicited through formal knowledge elicitation processes, with knowledge elicitation tools (e.g., modified number line instrument by Rajabally et al. [52]). The reduction from nine to three domain experts was due to availability and commitment.

Experts rated the links between each parent node and its child node on a numerical scale. A parent-child link represents the relationship between the parent and the child, and the impact of the parent on the child. The data for network calibration was obtained from the analysis of the link ratings. The ratings were first transformed into percentage link weights from which percentage probability values were derived for the CPTs of the child nodes. If, for example, a node, Q, has four parent nodes, A, B, C, and D, with parent-child links a, b, c, and d, with percentage link weights w_a, w_b, w_c, and w_d, respectively, the CPT for the child node, Q, is parameterized as shown in Table 2 (assuming that each node has two states (e.g., high (H) and low (L))). Thus, the CPT for each child node is built by summing the contributed percentage link weights of its parents [28], reflecting the type of interaction between parent and child nodes. The PPTs of nodes without parents (referred to as leaf nodes) were set to fair priors.

Different types of probabilistic interactions can exist between related nodes (parent(s) and child), such as conjunctive, compensatory, disjunctive, and inhibitory (see [53] for details). The interactions between the nodes in the LAP model are not modeled as compensatory because a low level on one of the parents of a child node would not be compensated for with a corresponding high level of one or the other parents. For example, in Fig. 3, a low on the node 7, “Proper Use of Equipment (PUE),” cannot be compensated for by a high on the node 4, “Ability To Work With Components (WWC)” or node 11, “Ability To Adapt (ADA),” all of which are parents (subproficiencies) of node 24, “Laboratory Abilities Skill (LAS).” The interactions are modeled as conjunctive in which a high-level status of the child node requires relatively high levels on all its parent nodes. This informs the adoption of the formula used in Table 2 for the quantification of the CPTs of the LAP model nodes.

6 EVIDENCE VARIABLE QUANTIZATION AND INSTANTIATION

The application of a BN model requires that evidence variables are instantiated for network update/inference. In this context, the evidence variables are first quantified, using evidential data, and then instantiated, using a fuzzification engine. Fig. 5 is a graphical representation of the evidential data extraction and evidence variable quantization and instantiation processes. The quantization process generates a numerical value, a score, for an evidence variable, which constitutes the input into the fuzzification engine, the output of which is a linguistic term corresponding to one of the states of the evidence variable. The main source of evidential data is the behavior log.
Behavior is said to be what a student “says” and “does” within the context of a given task [26]. The contents of the behavior log (consisting of atomic mouse click streams and key presses) are grouped into four evidential data sources:

1. Lab activity results and answers to lab questions,
2. built circuit schematics and simulation commands,
3. responses to pre- and postlab test items,
4. interaction events captured and logged by the events tracker/recorder agent in VEL.

Another source of evidential data is the students’ historical data that yields such information as students’ experiences using the VEL and their LSs.

6.1 Evidence Variable Quantization

The different evidential data source components of the behavior log are used to generate the necessary data for the quantization of the evidence variables. For every evidence variable, \( e_i \), \( i = 1 \) to \( k \) (number of evidence variables in the network), a set of data items, \( X_i = \{x_{ij}\}, j = 1 \) to \( m \), are required for its quantization. Rules for quantizing \( e_i \) using \( X_i \) were established with the help of the participating domain experts. The quantization process uses a reference data set and other relevant instructor specified data as the “gold standard.” The data extraction processes and quantization of the evidence variable, node 2 “CorrectPlacementOfComponents” (CPC), is described below, while that of nodes 8 “ApplyCorrectFormula” (ACF) and 22 “InterpreteExperimentalResult” (IER) are described in Appendix B (B1 and B2, respectively), available in the online supplemental material. The quantization of node 2 involves the comparison of a student’s built circuit with the reference circuit, for equivalence, by a CircuitComparator. Two electrical circuit networks are equivalent if they both contain the same set of components in type and value, and the manner in which the components are connected result in the same branch currents and branch voltages [54]. The circuits are represented in the system as netlists.

The quantization of node 2, CPC, necessitates that a student’s built circuit be compared to the reference circuit for topological equivalence, by the CircuitComparator. The topological comparison of two circuits requires the explicit comparison of the branch voltage and/or current characteristics of the circuits. The Kirchoff’s Current Law (KCL) equations of the reference and student circuits were compared, so that the comparison for topological equivalence was reduced to a comparison of two matrices. For this purpose, a directed graph, \( G \), without self-loops, is used to describe the circuit (network) and the KCL equations for the network derived from the graph [55]. The nodes and branches of the network make up the vertices and the edges of the graph, respectively. The orientations of the edges reflect the reference direction of the currents/voltages across the branches of the network. The concepts associated with the graph, \( G \), required for the derivation of the necessary matrices/equations for the characterization of the network represented by \( G \) include: Path, subgraph, loop, cutset, tree, and cotree [see [55] and [56] for details]. The KCL equations are generated using the Incident Matrix (IM), \( A_i \), which is derived directly from the graph, \( G \), for example, the network of Fig. 6, is described by the graph, \( G \), of Fig. 7.

In \( G \), the branches represent the branches of the network, so that \( a_1, b, c, d, e, f, g, \) and \( h \) represent \( v_{cc}, c_1, c_2, c_3, r_1, r_2, r_3, \) and \( r_4 \), respectively. \( G \) has \( n \) nodes and \( b \) branches corresponding to the \( n \) nodes and \( b \) branches of the network. The IM, \( A_i(1) \), is the \( n \times b \) matrix, \( A_i = [a_{ij}] \), where: \( a_{ij} = 1 \) if the branch \( j \) is incident at node \( i \) and the arrow is pointing away from node \( i \); \( a_{ij} = -1 \) if branch \( j \) is incident at node \( i \), and the arrow is pointing toward node \( i \); and \( a_{ij} = 0 \) if branch \( j \) is not incident at node \( i \). \( A_i = 0 \), where \( i \) is the \( b \times 1 \) column vector, representing the branch currents.

\( A_s = 0 \) is the set of \( n \) KCL equations characterizing the network, which are not linearly independent. The required \( n - 1 \) linearly independent KCL equations are obtained by transforming the IM, \( A_i \), to the Reduced Incident Matrix (RIM), \( A \), to derive the equation, \( Ai = 0 \), which is the required set of \( n - 1 \) independent KCL equations. The RIM is derived from the IM using the graph concepts of tree and cotree.
Placed correctly in this context, the trapezoidal function, of membership between 0 and 1, in particular type is built topologically correctly at least once). It maps each element, \( x_i \), for an evidence variable, \( A_i \), to a degree of membership between 0 and 1,

\[
A_i = \begin{bmatrix}
    a & b & c & d & e & f & g & h \\
    -1 & 0 & 0 & -1 & 0 & 0 & -1 & -1 \\
    1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
    2 & 0 & 0 & 1 & -1 & 1 & 0 & 0 \\
    3 & 0 & 0 & -1 & 0 & 0 & -1 & 0 \\
    4 & 0 & -1 & 1 & 0 & 0 & 0 & 1 
\end{bmatrix}
\]

The result of the comparison is used to generate the quantized value for the node, CPC, using (2), where \( n \) = number of different types of reference circuits, \( m \) = number of circuits built and simulated by student, \( x_i = \) number of circuits built and simulated by student that are topologically equivalent to reference circuit type \( i \), and \( q = \) number of \( x_i \)s not equal to zero \( (\mu_i \) is an indication that a circuit of a particular type is built topologically correctly at least once). CPC is quantized as:

\[
CPC = \frac{100}{n} \times q \left( \sum_{i=1}^{n} x_i \right) / m. \tag{2}
\]

### 6.2 Evidence Instantiation

The quantization process generates a numerical value (a score), for an evidence variable, which serves as input to the fuzzification engine. The fuzzification process transforms the score into one of the linguistic values (states) of the variable. The technicalities of the fuzzification process are documented in literature on fuzzy sets, logic, and systems such as [57]. In this context, the fuzzification process is based on the universe of discourse, \( X = \{0, 10\} \), the set of possible numerical scores for the evidence variables. The score, \( x_i \in X \), for an evidence variable, \( A_i \), is fuzzified by generating its membership values in \( T(A_i) = (t_1, t_2, t_3) \), the linguistic term set (states) of \( A_i \). In this context, the trapezoidal function, \( \mu(x_i) \), is the core of the fuzzification engine. The trapezoidal membership function is shown graphically in Appendix C, available in the online supplemental material, for node 2, which has three states: placedAllCorrectly, placedSomeCorrectly, and nonePlacedCorrectly. It maps each element, \( x_i \in X \), to a degree of membership between 0 and 1, in \( t_i \), where \( t_i \) is the set of ordered pair: \( \{t_i, \mu(x_i)\} | x_i \in X \). If, for example, the quantized value of \( CPC = 68 \), then \( \mu_{\text{placedSomeCorrectly}}(68) = 0.2 \) and \( \mu_{\text{placedAllCorrectly}}(68) = 0.8 \), based on which CPC will be instantiated to the state, placedAllCorrectly. A node is instantiated to the state for which the membership value of its score is greater.

### 7 Model Verification

Model verification is concerned with knowledge elicitation review, functional verification, and sensitivity analysis (SA). Review and refinement were inherent components of the knowledge elicitation process, so will not be addressed further.

#### 7.1 Functional Verification

The model was functionally verified by instantiating evidence nodes and checking whether the evidence was correctly propagated through the network. For example, when information about a student’s placement of components is received and it is ascertained that the student has placed all the components correctly, node 2 (CPC) is instantiated and clamped to the state “placedAllCorrectly.” Fig. 8 shows the relevant part of the LAP model before node 2 was instantiated, with the resulting updating of beliefs shown in Fig. 9.

As expected, the hypotheses, nodes 4, 15, and 16, that are influenced by node 2 have become more likely. The belief estimation for node 4, “abilityToWorkWithComponents,” rose from 50 to 64.3 percent. Also, there was a rise in the measure of belief, from 49.9 to 54.1 percent, in the ability to construct circuits and an increase in belief in node 16, “AbilityToModifyCircuit,” from 49.9 to 52.7 percent. There was a 1.00 percent increase in the measure of belief for performance.

#### 7.2 Sensitivity Analysis

SA is a technique for systematically investigating the effects of variations in inputs on a model’s output. Evidence-based sensitivity analysis (ESA) is used to detect and minimize the effects of poorly calibrated network or bias in knowledge elicitation by identifying the excessively influential variables. It estimates the extent to which the posterior probability of a target variable, \( X \), is changed when finding is entered at a node, \( A \), and the network updated. ESA was used to check the sensitivities of the performance node to the evidence nodes in the LAP model. The influence of node \( A \) on \( X \), referred to as the sensitivity of \( X \) to \( A \), or the contribution of \( A \) to the reduction of the uncertainty at \( X \), is
measured using Shannon’s measure of mutual information based on the entropy function [58], [59]. The total uncertainty reduction potential of $A$, with respect to $X$, is expressed as [58]:

$$S = H(X) - H(X|A)$$

$$= - \sum_{A} \sum_{X} P(x, a) \log(P(x, a)/P(x)P(a)),$$  \hspace{1em} (3)

where $H(X)$ is the initial uncertainty in $X$, before receiving the evidence, $A$, and updating the network, and $H(X|A)$ is the average residual uncertainty in $X$, summed over all possible values of $A$.

The sensitivity of the target node, node 40, “performance” ($PF$), to all the 23 indices (evidence variables) in the LAP model, and the Memory/Feedback node, node 35, “PreviousPerformance” ($PP$), is depicted graphically in Fig. 10. Nodes 35, “PP,” and 8, “ACF” can be seen to have excessive influence on performance. The influence of node 8 was moderated through the adjustment of its contributing weights in the parameters of its children, in consultation with the domain experts. The result of moderation is highlighted in Fig. 11. The moderation caused a desired increase in the influences of nodes 38, “ConceptualUnderstanding” (CLU), and 36, “FactualKnowledge” (FKN). Experts agreed that the level of influence of node 35 was acceptable so its influence did not require moderation. Also, the levels of influence of nodes 38 and 36 (Fig. 11) was considered desirable because it is in line with the domain experts’ opinions that they should impact equally and more significantly on performance.

### 7.3 Model Evaluation

The aim of evaluation is to verify the reliability and validity of a model’s assessment outcomes.

#### 7.3.1 Reliability

Reliability is concerned with the extent to which an assessment, if repeated, would give the same results [60]. It is an estimation of the consistency or repeatability of an assessment. An assessment is considered reliable if the assessment outcome for two students with similar abilities/aptitude is similar, for the same assessment. An assessment is considered reliable if repeated, would give the same results [60].

Reliability is concerned with the extent to which an assessment, if repeated, would give the same results [60]. An assessment is considered reliable if repeated, would give the same results [60].

Another method of reliability evaluation is the derivation of predictive accuracy measure, based on the use of training and test data sets [33], [61]. This method is most appropriate for models induced from data [33]. It is, therefore, not applicable to LAP model reliability testing. To evaluate the reliability of the LAP model, simulated students evaluation approach was adopted. This approach has been used by other researchers for the evaluation of assessment models and systems including [34], [37], [62], [63]. In this context, the construction of the simulated students was based on [68] and space restrictions constrain the importation of the description of the full details of the creation process.

Simulated students are models of human learning and have been categorized based on their granularity: Large and fine-grained [64]. The granularity of a simulated student refers to the amount of detail in its representation [64]. Large-grained simulated student models, with known states of performance indices and scores for the performance criteria, and therefore known overall performance scores, were constructed and used for the evaluation. The overall performance score is arrived at by normalizing the sum of the contributing link weights of the known states of the performances indices with respect to the maximum link weights of the nodes. The LAP model was ran for each of the simulated students, with the evidence variables instantiated to their known states, to obtain the model’s performance belief estimates for the students. Fig. 12 is a graphical depiction of the known performance scores compared to the estimated measures of belief in performance.

There was a strong positive correlation between the students’ known performance scores and the estimated beliefs ($r = 0.87685, p = 0.0000$). For reliability, “an assessment with a correlation coefficient that is less than 0.70 is generally not considered suitable for individual student evaluations” [37]. A high correlation between scores obtained from two assessments indicates that the assessment is reliable [65]. Where the known performance scores had mean, $\mu = 51.79$, and standard deviation, $\sigma = 18.11$, the estimated measures of belief in performance had $\mu = 50.23$ and standard deviation $\sigma = 14.46$.

#### 7.3.2 Validation

Validation relates to the interpretation of the outcomes of an assessment and the evidence to support a specific
interpretation [66]. The different types of validation checks include face, content, criterion, and construct validity [37], [60]. Face validation checks if, on the face value, an assessment process and the outcome appear relevant to the assessed construct. Content validation checks the assessment tasks for relevance and representativeness of the domain. Content validity is often established by agreement in the judgements of domain experts. Criterion validity is of two types: Predictive and concurrent. Predictive validity check is deemed most appropriate for assessment models that predict future performance, such as aptitude tests [37]. Concurrent validity often entails the comparison of the results of an assessment, based on a particular model or scheme, to the results of the gold standard, both of which reflect the same construct. Construct validity checks the extent to which an assessment model measures the specific construct for which it is designed.

In this context, domain experts established the face validity of the assessment outcomes of the LAP model, having been key participants in the design and parameterization of the model. Content validity is assumed implicit because domain experts’ were actively involved in the design of the assessment tasks. Predictive validation is not relevant to this work. The focus was on concurrent validity, which checks the accuracy of assessment outcomes, for a set of assessment scenarios.

The LAP model was evaluated using the virtual observation technique. The validation process involved a team of four assessors. The assessors, rather than physically observe the students themselves, used the “observations” (students’ behavior logs) made and recorded by the events tracker/recorder component of the VEL, while students were undertaking lab activities in the virtual laboratory environment, to assess students’ lab work performance. They used the four evidential data source instruments that make up the behavior log to score a set of performance indicators, based on their respective set of scoring criteria. The assessors were, a priori, familiarized with the research work, the nature and content of the assessment instruments, and the assessment process. The assessors worked individually in the same physical space, with minimal discussions between them to avoid any undue influence on each other. One of the researchers facilitated the assessment process by answering assessors’ questions and clarifying relevant unclear issues. The assessors each assessed the same set of two lab activities for each of the 52 students, with respect to a set of performance indicators and their related criteria. The focus was on seven of the nine performance indicators in the LAP model, for the purposes of validation, considering time and cost.

As part of the validation process, it was sought to answer the following questions: Q1—are the assessors consistent in their assessment?: Q2—to what extent do the assessors agree with each other?: and Q3—to what extent do the assessor-assigned students’ performance scores agree with the LAP model estimated beliefs in students’ performance. To verify assessor consistency, each assessor assessed one of the laboratory activities twice for each student, using the same set of assessment instruments. Scores from the two independent assessments were then compared. At $\alpha = 0.05$ and two tailed significance tests, the assessors were statistically significantly consistent in their assessment, with respect to the assigned scores for the performance indicators. Assessors $A1$, $A2$, and $A3$ were highly consistent ($0.50878 \leq r \leq 0.84319$, and $0.00000 \leq p \leq 0.000118$). Assessor $A4$ was moderately consistent ($0.33141 \leq r \leq 0.67701$, and $0.00000 \leq p \leq 0.024006$), providing a positive answer to Q1.

Next, interassessor scores correlation was verified, to answer question Q2. Table 3 lists the interassessor scores correlation factors, $r$, for one of the lab activities, with respect to a set of performance indicators. $A1$ was in agreement with $A2$, on average, 75.63 percent of the time, with $A3$, 64.88 percent, and with $A4$, 33.94 percent. $A2$ was in agreement with $A3$ 67.86 percent of the time, and with $A4$, 37.77 percent. $A3$ agreed with $A4$ only 41.99 percent of the time, on average. Assessors $A1$, $A2$, and $A3$ agreed with each other more than 64 percent of the time. Assessor $A4$ agreed the least with all other assessors.

To answer question Q3, the LAP model was used to assess the same set of laboratory activities as the assessors and the assessment outcomes compared to those of the assessors. Evidence variables were quantized from the same set of assessment instruments used by the assessors and then instantiated into findings or hard evidence. The findings were entered into the network and the target nodes queried. The estimated beliefs for the target nodes were recorded and compared to the scores assigned by the assessors. The human assessors had access to the same set of assessment instruments and data as the LAP model, and used the same set of indicator criteria. Hence, the human performance assessment outcomes were used as the “gold standard.”

![Fig. 12. Test for reliability.](image)

**TABLE 3**

<table>
<thead>
<tr>
<th>Performance Indicator</th>
<th>$A1$</th>
<th>$A2$</th>
<th>$A3$</th>
<th>$A4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$pse$</td>
<td>0.79508</td>
<td>0.60305</td>
<td>0.40658</td>
<td>0.74067</td>
</tr>
<tr>
<td>$src$</td>
<td>0.62128</td>
<td>0.74617</td>
<td>0.31646</td>
<td>0.64266</td>
</tr>
<tr>
<td>$wve$</td>
<td>0.85274</td>
<td>0.69423</td>
<td>0.44976</td>
<td>0.69267</td>
</tr>
<tr>
<td>$sae$</td>
<td>0.76539</td>
<td>0.67283</td>
<td>0.41472</td>
<td>0.74912</td>
</tr>
<tr>
<td>$ise$</td>
<td>0.77081</td>
<td>0.50095</td>
<td>0.39649</td>
<td>0.61171</td>
</tr>
<tr>
<td>$mac$</td>
<td>0.73432</td>
<td>0.52202</td>
<td>0.16400</td>
<td>0.60400</td>
</tr>
<tr>
<td>$ege$</td>
<td>0.79147</td>
<td>0.79111</td>
<td>0.23181</td>
<td>0.70951</td>
</tr>
</tbody>
</table>
Strong positive statistically significant correlations existed between the human performance assessment outcomes and the LAP model performance belief estimation outcomes ($PUE: r = 0.77795; WW C: r = 0.88982; PF: r = 0.80157$). The $p$-values for all three indicators were 0.0000. The average human assessor score compared to the model estimated belief, for the performance indicators, $PUE$, and $WWC$, and $PF$, are shown graphically in Figs. 13, 14, and 15, respectively.

8 MODEL APPLICATION

This section highlights the possible application processes of the model. The idea behind the application of a BN model is to obtain estimates of certainties for hypotheses or variables that are unobservable (or observable at an unreasonable cost), based on findings for the observable variables. In this work, the model application is aimed at inferring, within the VEL environment:

1. What a student is actually able to do in terms of specific laboratory abilities/skills (competencies);
2. a student’s grasp of the concept addressed by a laboratory activity;
3. a student’s overall performance.

Inference (belief estimation) is based on the findings generated from the data derived from student’s behavior in the course of undertaking laboratory activities within the environment. The application of the model is based on entering the findings at the evidence nodes, updating the network, and reading off the estimated beliefs of the target nodes. Network update calculates the posterior probabilities of the target nodes.

There are 23 evidence variables in the LAP model. It may not be possible to generate findings for all the evidence variables for every laboratory activity because “it may not be necessary to assess everything all of the time” [12]. Each laboratory activity does not necessarily have to generate performance data to evidence all the observable nodes and the decision to enter all or some of the findings, at any assessment instance, depends on the instructor. For example, if a laboratory activity does not involve calculations, then nodes 8 ($ACF$) and 9 ($UCV$) (see Fig. 3) cannot be evidenced. Also, the instructor may seek to evidence a specific set of variables to make inference of a specific node. To do this, the instructor may choose the nodes to evidence and the target nodes to query, using the appropriate options in the instructor interface GUI, which has options to enable the instructor turn the evidencing of variables and reasoning about variables on and off. For example, the instructor may design an activity that will specifically require students to adapt, then turn off the evidencing of other nodes, except those required to infer a student’s ability to adapt. Also, the instructor may wish to focus on a student’s grasp of the addressed concept(s) by turning off the evidencing of other nodes, except nodes 8 ($ACF$), 9 ($UCV$), 19 ($FOI$), 20 ($RDG$), 25 ($CUN$), and 27 ($FKN$) and then querying nodes 37 ($KNC$) and 39 ($UNC$). If no variables are turned off, findings are generated for all evidence nodes, where possible, and reasoning generalized.

The LAP model can be used to assess each lab activity instantaneously, which in this context, entails: The use of the BN model without the memory/feedback node and the update of the network, in the light of new evidence, after the previously estimated beliefs, based on previous evidence, are cleared from the network and the network reset to its original state with fair priors. New findings are then entered and inference made to obtain estimates of certainties for target variables. In this case, the network serves as an inquiry-based assessment tool to possibly diagnose areas of improvement. Alternatively, assessment could be based on the combined effects of past performance outcomes and present activity, referred to, in this context, as cumulative assessment. Cumulative assessment gives the instructor an integrated picture of a student’s performance, learning, and developmental progress over time. This is achieved from two perspectives: Nonretractions of previous findings entered into the network; and the activation of the memory/feedback node.
9 DISCUSSION AND CONCLUSION

9.1 Discussion

LAP model construction was conclusively undertaken with the committed participation of three domain experts, down from an initial nine. The experts consisting of a Principal Lecturer, a Senior Lecturer, and a Reader were from the Electrical and Computer/Electronic Engineering (ECE/EEE) Departments of two different universities. Each expert had over 10 years classroom/lab experience. The assessors that participated in the evaluation of the model were different from the experts that helped formulate the model. The model was evaluated in an engineering faculty (with large class sizes and resource constraints) in a country in developmental transition. The assessors were unfamiliar with the VEL, the assessment tool, and the virtual observation-based assessment scheme. The introductory meetings, demonstrations, and briefing sessions facilitated their understanding and appreciation of the research work, which increased their motivation. After the first assessment session, assessors had time to reflect on the assessment process and their experiences, the noticeable effect of which was increased speed of assessment, shorter pause durations, and fewer questions to the researcher. At the end of the assessment process, the assessors’ feedback on the VEL, the assessment tool, and the assessment methodology were encouraging, indicating that the VEL, the assessment tool, and scheme had been well received. The assessors acknowledged the potential of the assessment tool and scheme to avail the instructor of useful information about what students are actually able to do, their lab work schema, and knowledge and understanding of the concept addressed by a laboratory activity.

The virtual-observation assessment scheme has the potential to offer the same benefits as the physical-observation scheme. This potential is further strengthened by the strong positive correlations that exist between the assessors’ scores and the belief estimates by the LAP model. Though the agreement between the assessors was not always good, that is not a problem in itself but more the fact that (at the time the study was being conducted) it was not possible, for cost and logistical reasons, to collect a larger data set (i.e., involving more than four assessors). While assessor 4 is likely to have been a glitch, it may have represented a genuine “trend.” The outcome of the comparison shows strong positive correlation despite assessor 4’s contribution, which is encouraging and motivates further investigation as part of future work. The directions for future work have been highlighted in the next section.

The main contributions of this work are threefold. First is the design, testing, and validation of the LAP model for the performance assessment of students’ lab work in a VEL environment, from a holistic perspective. The potential of the model as an assessment tool has been demonstrated. Second is the demonstration of the feasibility of the performance assessment of students’ lab work from their observed behavior in a VEL environment, based on the concept of virtual-observation, as an alternative to the written evidence technique often extended from the traditional laboratory to the virtual laboratory environment. Third is the demonstration of the feasibility of the virtual-observation technique as an assessment model validation technique for VLEs.

9.2 Conclusion

The important role accorded to lab work in the UE course reflects its value as an instruction/learning tool. Assessment drives learning and the assessment of students’ lab work performance is challenging. A review of literature highlighted the lack of meaningful assessment tools for virtual laboratory environments in engineering education. Literature review also highlighted the shortcomings of traditional student lab work assessment practices, among engineering faculties. There are also the problems of time demands, bias, inconsistency, and increasing student numbers which make the traditional lab work assessment scheme impractical. The physical-observation assessment scheme that was proposed as a possible alternative to the written evidence technique is costly (time, space, human resource, logistics), and difficult to apply to large class sizes. In view of these and the increasing adoption of virtual lab in engineering, the LAP model (described in this paper) was designed for the performance assessment of students’ lab work in a VEL environment based on the virtual-observation technique. The model has been described in detail and its evaluation results presented. The validation results are encouraging and future work will encompass the generation of larger data sets by deploying the model into practice with real students and real laboratory scenarios, and the involvement of more assessors. The assessment model and assessment scheme have the potential to reduce the cost and burden of marking students’ lab work, promote fair, consistent and timely assessment, and ease the handling of larger class sizes. The LAP model was evaluated using the virtual observation technique to demonstrate the potential of the LAP model as a performance assessment tool; feasibility of the performance assessment of students’ lab work from their “observed” behavior in a VEL; and the feasibility of the virtual observation technique as a validation evaluation technique for assessment models.

ACKNOWLEDGMENTS

This work was supported in part by the Schlumberger Foundation under its Faculty For The Future scholarship program aimed at encouraging women in science, engineering, and technology in their pursuit for academic excellence.

REFERENCES


Ifeyinwa E. Achumba (formerly Ifeyinwa E. Chika) received the BSc degree in computer science from the University of Nigeria, Nsukka, and the MSc degree in computer engineering from the Federal University of Technology, Owerri, Nigeria. She is currently a research student in the Electronic and Computer Engineering Department of the University of Portsmouth, United Kingdom. The research degree program is funded by the Schlumberger Foundation under its Faculty for the Future Scholarship program aimed at encouraging women in sciences, technology, and engineering in their pursuit of academic excellence. She is a member of the IEEE.

Djamel Azzi received the Ingénieur D’État degree in electronics from the Université de Constantine, Algeria, in 1988, and the MSc degree in control technology in 1991 and the PhD degree in control engineering in 1993 from the University of Manchester Institute of Science and Technology, United Kingdom. He is currently a principal lecturer in the Electronic and Computer Engineering Department at the University of Portsmouth, United Kingdom. His research interests include all aspects of AI but in particular its application to robotics, wireless sensor networks, and assisted living.

Victor L. Dunn, received the MSc degree in microelectronics systems. He is currently a senior lecturer in digital electronics and microprocessor systems in the Electronic and Computer Engineering Department at the University of Portsmouth, United Kingdom. He was previously with GEC Marconi. He is the leader for the award of student prizes and the IEEE students’ liaison officer. He is a member of the IET and fellow of the HEA.

Gloria A. Chukwudebe received the BEng, MEng, and PhD degrees in electronic and computer engineering. She is currently a reader in the Electrical and Electronic Engineering Department at the Federal University of Technology, Owerri (FUTO), Nigeria. She has several publications and has presented many papers at national and international conferences. She has also been an invited guest speaker for several organizations in Nigeria. She teaches courses in electronic and computer engineering up to the postgraduate level. She was the head of the Department of Electrical/Electronic Engineering at FUTO from 1999-2005. She is an active IEEE volunteer.