

# Improved computation of individual ZPD in a distance learning system.

Sara De Cristofano, Maria De Marsico, Andrea Sterbini

Department of Computer Science  
Sapienza University of Rome  
Rome, Italy

[sara.decrstofano@gmail.com](mailto:sara.decrstofano@gmail.com), {demarsico,  
sterbini}@di.uniroma1.it

Marco Temperini

Department of Computer, Control, and Management  
Engineering Antonio Ruberti  
Sapienza University of Rome.  
Rome, Italy  
marte@dis.uniroma1.it

**Abstract**—This paper builds upon theoretical studies in the field of social constructivism. Lev Vygotsky is considered one of the greatest representatives of this research line, with his theory of the Zone of Proximal Development (ZPD). Our work aims at integrating this concept in the practice of a computer-assisted learning system. For each learner, the system stores a model summarizing the current Student Knowledge (SK). Each educational activity is specified through the deployed content, the skills required to tackle it, and those acquired, and is further annotated by the effort estimated for the task. The latter may change from one student to another, given the already achieved competence. A suitable weighting of the robustness (certainty) of student's skills, stored in SK, and their combination are used to verify the inclusion of a learning activity in the student's ZPD. With respect to our previous work, the algorithm for the calculation of the ZPD of the individual student has been optimized, by enhancing the certainty weighting policy, and a graphical display of the ZPD has been added. Thanks to the latter, the student can get a clear vision of the learning paths that he/she can presently tackle. This both facilitates the educational process, and helps developing the metacognitive ability self-assessment.

**Keywords**—Zone of Proximal Development; Student Model; Reachability of a Learning Activity

## I. INTRODUCTION

The research presented in this paper stems from theoretical studies in psychology, and particularly in the pedagogy field, known by the name of social constructivism. Lev Vygotsky is considered one of the most influential representatives of this research line, with his theory of the Zone of Proximal Development (ZPD) [1]. In a few words, this cognitive region includes those learning activities that the student can safely tackle, especially if supported by group collaboration. It is in the middle between two other cognitive zones. The zone of Autonomous Problem Solving includes concepts, knowledge and skills that are firmly possessed by the learner, such that it is possible to autonomously tackle the solution of problems therein. In a symmetric way, the zone of Unreachable Problem Solving includes concepts, knowledge and skills that are not possessed at all, or only in a too weak way, such that tackling related problems, though with the help of teachers and peers, would only frustrate the student. This theory especially

underlines the value of group activities, and inherently postulates *social* achievements. Our previous work started by extending the concept of ZPD to an individual dimension first [2][3]. The research efforts aimed at including the concept into a real system: the core problem is to find a possibly objective way to estimate the *cognitive distance* of a learning activity from the student's *cognitive state* by suitably weighting the already achieved abilities [4]. The research has then proceeded towards the definition of a group ZPD [5].

The overall theory developed is based on the assessments carried out during the educational process of individual students. For each learner, the results of such assessments are used to build and store an evolving model summarizing the Student Knowledge (*SK*). The model associates, to each knowledge chunk (*skill*), the estimated degree of competence for the student (*certainty*). A minimum value of certainty is requested to include a skill within the *SK*. Each educational/learning activity (*la*) is specified by the deployed content, but also by the skills required (*la.P*) to tackle it, and those acquired (*la.A*). Furthermore, the creator of a specific activity mentions the expected effort estimated for the task. The idea which this work builds upon, is that if the student already owns some elements in *la.A*, then the actual effort can decrease, while it increases if the student lacks a sufficient certainty in some elements in *la.P* [4]. In practice, considerations about the owned skills and the respective certainty are used to compute the weight of each skill in computing the *reachability* of a learning activity. Reachability is the parameter that determines the inclusion of a learning activity into a student's ZPD.

With respect to our previous work, the algorithm for estimating the inclusion of a learning activity into the ZPD of an individual student has been optimized, by adopting a more suitable computation of skill weights to better estimate the so called *daring threshold*. This threshold is used in our previous work to determine the dynamic border of the ZPD, which changes depending on the learner, on the activity, and of course on time. In practice, it regulates the amount of effort that is reasonable (pedagogically viable) to require from the student in tackling an educational activity. In other words, it helps determining which learning paths can be fruitfully tackled.

Both versions of the computation of certainty weight have been tested to determine which one provides results expectedly closer to real situations. We carried out some tests in vitro to compare the newly introduced weighting policy. The achieved results are quite satisfying and testify that the ZPD (its width and composition) is a good index to consider to make students tackle activities in the most productive way.

As a further contribution of the present work, it introduces a graphical display of the graph of the activities within the student ZPD. This allows to provide the student with a clear vision of the learning paths that he/she can currently tackle, according to the achieved cognitive state. Besides facilitating the educational process, this can also be considered a useful tool to develop the metacognitive ability of self-assessment.

## II. RELATED WORK

At present, the literature related to distance learning is extremely huge, and addresses from many perspectives the multi-faceted problems related to the design of pedagogically effective applications. Therefore, attempting a comprehensive discussion of the related work about this topic is quite unfeasible. It is better appropriate to just mention some specific interest points and some examples of works dealing with the main issues. Furthermore, as the best of our knowledge, there is no other approach at present in literature that aims at integrating the theory of ZPD in a real e-learning system.

A first critical element in the design of distance education is personalization of learning activities, supported by an effective learner profiling. The design and implementation of tools and strategies to support the construction, maintenance and delivery of adaptive courses is among the hottest topics of the research about distance learning. Designing educational applications, especially when targeted at distance learning activities, raises a number of specific challenges, including an adaptive sequencing of tasks, most of all in large-scale web-based education, as discussed by Brusilowsky and Vassileva in [6], that must be supported by a suitable student model, as investigated by Brusilovsky and Millan in [7]. The latter is often based on the preliminary determination of individual learning styles, for example the one used by Graf and Kinshuk in [8]. Student Cognitive State and Learning Style are also used by Limongelli *et al.* in [9] to build a system capable of providing Educational Hypermedia with adaptation and personalization. A complementary problem is to provide course content creators with effective tools to support the design of adaptive learning, as the ones proposed by De Bra *et al.* in [10]. Silius and Tervakari [11] use the specific term *pedagogical usability* for the specific property of tools, content, interface, and tasks of e-learning systems to support different learners to learn in different educational contexts according to selected pedagogical objectives.

A second element, which is of paramount importance and may lack in distance education, is a suitable social interaction with the teacher, and most of all with peers. According to consolidated guidelines of modern pedagogy, students must be “immersed” in a comprehensive framework fostering social activities. Collaborative learning is universally recognized as a winning methodology to allow the development of both

cognitive and meta-cognitive abilities, such as critical thinking. Kreijns *et al.* also underline that it can support better retention and deepening of knowledge over time [12]. Collaborative activities better prepare the learners for real-life team-based working, by sharing their experience, and combining their skills, as investigated by Cheng and Ku in [13]. According to these considerations, the system E-MEMORAe2.0 presented by Leblanc e Abel in [14] is an example of an e-learning environment implementing Organizational Learning.

## III. VYGOTSKIJ’S ZONE OF PROXIMAL DEVELOPMENT

The elements presented above converge in Vygotskij’s guidelines. One of the most relevant lines from Vygotskij’s research [15] aims at demonstrating that cooperation is a critical factor in setting the basis for the individual development. Since from childhood, any activity implying some kind of interaction, either targeted or even casual, activates and spurs important cognitive processes. Development stems from the interaction between individuals and the environment. Vygotskij looks at learning as a result of exchanges of learning results occurring in social interaction, and especially between a more experienced person and a less competent one. In the model by Vygotskij, the direction of learning is from outside to inside, as in traditional theories, except that the knowledge interiorization happens through the social “co-construction” (social learning), by proceeding through a progressive transfer of the exterior social activity to the interior control. Once knowledge and processes are interiorized, the learner will be able to proceed autonomously.

The importance attributed to the social cognitive development is reflected in the concept of “Zone of Proximal (or Potential) Development” (ZPD). The interaction and peer support exchange triggers the potential for growth. While the presence of the “others” was formerly a real co-presence, it has to become also virtual nowadays, relying on present advances in technology. Concrete growth can occur only in the ZPD, which is characterized by the distance between the actual development level, as it is determined by the autonomous problem-solving (APS), and the level of potential development, as determined by the problem-solving under teacher’s guidance or in collaboration with peers. In practice, it is defined as the area of learning that stays between what one can do individually, in terms of knowledge, skills, and abilities, and what one cannot do not even if helped. From a pedagogical points of view, once the learner has consolidated a region of Autonomous and independent Problem Solving (APS), it is useless to further suggest exercises related to the same level of difficulty. On the other hand, it may result even worse to suggest exercises which are completely out of reach for the learner, despite any possible support (Unreachable Problem Solving – UPS). This will only cause frustration and demotivation. The right zone of complexity includes activities that the learner can carry out, according to owned competence and/or with a the support of collaboration with companions.

The aim of our ongoing work is to integrate into an e-learning system the appropriate model elements and procedures to dynamically determine when a learning activity lies inside the ZPD area either of a single student, or of a working group.

#### IV. BASICS OF THE LEARNING MODELLING FRAMEWORK

For reader's sake, before introducing the present proposal, it is necessary to summarize the basic definitions related to the framework integrated with ZPD evaluation, and from which the present refinements stem. The Student Model, namely  $SM(l)$ , is used to model the single *Learner* (individual)  $l$ , and stores information about the *learning style* and the *cognitive state* of each student. The latter component, namely  $SK(l)$ , records at any moment the current state of skills for student  $l$ ; it is upgraded as the student tackles learning activities, and is the model part that is relevant for the present work. Each skill represents an ability/knowledge related to a certain learning domain. A learning activity is expected to trigger the acquisition of certain skills, while it may require a set of prerequisite skills. A skill is defined as a predicate  $S()$  whose arguments include: a *main concept* (a conventional name); an integer value (*level*) and a *keyword*, expressing the cognitive level to which the concept is possessed according to Bloom's taxonomy [16]; an optional *matching concept* (possibly multiple), to which the achievement of the main one is related; and a *context*, stating the disciplinary context for the occurrence of the concept(s):

$$S(\text{concept}, k, \text{keyword}, \text{matching\_concept}(s), \text{context})$$

$SK(l)$  is a set of pairs representing both the skills possessed and an evaluation of their firmness (*certainty*), e.g., the *confidence* assigned individually to the achievements:

$$SK(l) = \{ \langle s_i, c_i \rangle, \dots, \langle s_{n_i}, c_{n_i} \rangle \}$$

The certainty  $c_i$  for a skill  $s_i$  is a number  $c \in [0..1]$  (an higher certainty corresponds to greater confidence) and is computed/updated following the assessment activities undertaken by the learner during the course. The first time a learner  $l$  acquires a skill, the couple  $\langle s, C_{ENTRY} \rangle$  is added to  $SK(l)$ , where  $C_{ENTRY}$  is a default starting confidence. The certainty is either increased after each further successful assessment for  $s$ , or decreased after each unsuccessful one. In this way, the state of skills for  $l$  is a dynamic element that always shows the current evaluation of certainty for the achieved skills. When the certainty for  $s$  decreases below a level  $C_{DEMOTED}$  the couple  $\langle s, c \rangle$  is deleted from  $SK(l)$ : further activities will be needed to acquire it back; on the other hand, when in  $\langle s, c \rangle$  exceeds a conventional value  $C_{PROMOTE}$ , the  $s$  is given as firmly acquired, and no further specific assessment is required for it.

A *learning activity* is designed to be carried out either individually, or as a social-collaborative group work. The components of a learning activity  $la$  are the following:

- *la.Content* – a collection of learning material, possibly entailing the use of a supporting software platform;
- *la.A* – Acquisition: a set of skills the student is expected to achieve after tackling the  $la$ ;
- *la.P* – Prerequisites: a set of skills that are required to fruitfully tackle the  $la$ ;
- *la.Effort* – an estimate of the cognitive load associated to tackling the  $la$ .

Successful completion of  $la$  by student  $l$  will entail the integration of the  $la.A$  set of skills into  $SK(l)$ , each with certainty value  $C_{ENTRY}$ ; for skills already in  $SK(l)$ , the

corresponding assessment will cause an increase (or a decrease) of certainty.

A repository  $R$  of learning activities is a set of *las*, available for building courses related to a specific Knowledge Domain, that in turn is handled as a set skills.

A *learning path* is defined as a set  $LP = \{la_i\}_{i \in \{1..n\}}$ .

For a  $LP$  we can identify the overall requirements  $LP.A$ , and the overall requirements  $LP.P$  as

$$LP.A = \cup_{i \in \{1..n\}} la_i.A \quad LP.P = \cup_{i \in \{1..n\}} la_i.P \setminus LP.A$$

and the overall effort imposed by  $LP$  on a learner as

$$LP.Effort = \sum_{i \in \{1..n\}} la_i.effort$$

This estimation of the effort entailed by a learning path, and more generally by a path of activities required to reach a certain skill, is based on the expected effort estimation associated to each activity. As discussed in the following, such estimation might be made more accurate on a per student basis.

For a student  $l$  to be able to profitably access an activity  $la$ , all skills in  $la.P$  should be present in  $SK(l)$ , meaning the student currently achieves them with a certainty of at least  $C_{ENTRY}$ .

The *s-projection* of the set  $SK(l) = \{ \langle s_i, c_i \rangle, \dots, \langle s_{n_i}, c_{n_i} \rangle \}$ , is defined as the set of skills appearing in the set pairs:

$$s\text{-proj}(SK(l)) = \{s_i, \text{ with } \langle s_i, c_i \rangle \in SK(l)\} = \{s_1, \dots, s_{n_i}\}$$

The definition of learning activity implicitly allows to define a *relation of derivation* (*propedeuticy*): given two learning activities  $\langle la', la'' \rangle$ , if  $la'.A \cap la''.P \neq \emptyset$ , some skills needed to tackle  $la''$  are acquired through  $la'$  and  $la'$  precedes  $la''$ . This induces a *relation of partial order* in a repository  $R$  of learning activities related to a specific topic, which allows depicting it as a graph. Every course is a subset (subgraph) of  $R$ . More details on framework definition can be found in [4].

In current deployment policies, the course is linearized according to the relation of derivation. Such linearization is possibly not unique. Letting the learner choose the order of *las* to take, with the only condition to comply with prerequisites, may increase independence and motivation in attending the course. In practice, the choice of the next learning activity should be limited only by the current possibility to tackle it, computed according to the current state of skills  $SK(l)$ . Vygotskij's theory is a rich source of inspiration to design the student-system co-evolution pattern just depicted, and entails the concepts needed to support a truly social-collaborative approach to taking learning paths.

A learning path/course  $LP$  entails a subset of the pertaining knowledge domain, denoted as  $KD(LP)$ . Given a learner  $l$ , working on a course  $LP$ , we can define some significant cognitive areas related to student's learning state according to Vygotskij's theory. Of course, we are especially interested in devising an operative definition of the ZPD. In order for  $l$  to profitably tackle a learning activity  $la$  in  $LP$ , all skills in  $la.A$  should be contained in the individual student ZPD.

Autonomous Problem Solving (APS) is the area of firm knowledge, in particular including skills related to the course:

$$APS(l) = \{s \in KD(LP) \mid \langle s, C_{PROMOTE} \rangle \in SK(l)\}$$

It is possible to observe that  $APS(l) \subseteq s\text{-proj}(SK(l))$ .

Proceeding from the available definitions of  $SK(l)$  and  $APS(l)$ , a straightforward definition of the Zone of Proximal Development (ZPD) of the learner  $l$  in the course  $LP$ , is

$$ZPD(l) = s\text{-proj}(SK(l)) \setminus APS(l)$$

Among the skills entailed by the course, this is an area where the learner can be expected to be able to go, with the help of teacher and peers, and strengthen skills already owned.

According with this definition of ZPD, the area of Unreachable Problem Solving (UPS), i.e. the subset of the course activities, that it is not pedagogically sound for the learner to tackle, given her/his present state of knowledge, can be determined as a complement:

$$UPS(l) = LP.A \setminus (APS(l) \cup ZPD(l))$$

It is immediate to consider that ZPD can be defined in a more challenging way, which may better stimulate the student. In fact, even skills not currently present in  $SK(l)$  might be profitably tackled, depending on their *cognitive distance*.

#### V. DISTANCE OF A SKILL FROM THE STUDENT COGNITIVE STATE

Given a learner  $l$  and a skill  $s$  still outside  $SK(l)$ , it is possible to use the above mentioned relation of partial order among learning activities to identify a set of possible (sub)learning paths  $G$ , contained in a course  $LP$ , that take to possibly acquire  $s$ , and that are traversable from the current state of skills:

$$\begin{aligned} Reach(s, SK(l), LP) &= \\ &= \{G = \{la_i\}_{i \in \{1 \dots n\}} \subseteq LP \mid s \in la_n.G.A \wedge la_i.A \subseteq s\text{-proj}(SK(l)) \wedge \\ &\quad G.P \subseteq s\text{-proj}(SK(l)) \cup G.A\} \end{aligned}$$

Notice that the last condition relating  $G.P$  to  $G.A$  accounts for the possibility that the prerequisites of some  $la_i \in G$  might be acquired through a preceding  $la_j \in G$ .

The distance of  $s$  from the present  $SK(l)$  is defined as

$$D(s, s\text{-proj}(SK(l)), LP) = \min_{G \in Reach(s, SK(l), LP)} G.Effort$$

The subset of skills, already in  $SK(l)$ , that are necessary to reach  $s$  along the identified minimal-effort path  $G^*$  in  $L, P$  is defined as the *support set* to reach  $s$  and denoted as

$$Support(s, SK(l), LP) = G^*.P \cap s\text{-proj}(SK(l)).$$

The problem to solve is how to determine  $G^*$ . This can be considered as a problem of finding a minimal cost path in a graph. It is first necessary to identify the characteristics of such a graph, as induced by the propedeuticity relation. Figure 1 shows an example, where nodes represent skills, and arcs are learning activities and are labeled by the estimated effort stored in their definition within the repository (for simplify the presentation, we avoid the further label with activity identity). The relation between arcs and nodes is as follows: a node that represents a skill  $s_i$  has as many incoming edges as there are learning activities  $la_j$  for which  $s_i$  appears in  $la_j.A$ , i.e., activities by which the skill can be acquired, and as many outgoing edges as there are learning activities  $la_k$  that require the skill  $s_i$ , i.e., such that  $s_i$  is included in  $la_k.P$ . This kind of graph has a peculiarity: it is possible that reaching a skill requires dealing with more activities, because it can happen that the student has to acquire further skills before the one in a node.

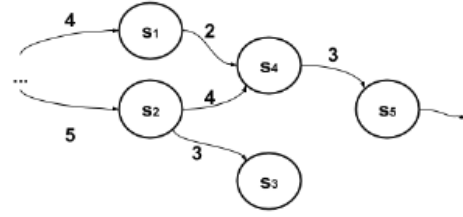


Figure 1. Example of graph with skills and activities.

Taking the example of the previous Figure 1, reaching the skill  $s_5$  requires acquiring both the arcs that lead to it, in this case the sub-path that provides for the acquisition of skill  $s_1$  and the other one that provides for the acquisition of skills  $s_2$ , and this must be taken into account to choose the final path of minimum cost. The graph is then of the *and-or* type: a type of graph that is often used to represent logical processes, which defines a relation of implication, conjunction or disjunction between its arches. In particular, the *implication* is represented by an edge connecting a *premise* node to a *consequence* node; the *conjunction* is a relationship between multiple nodes represented by arcs (*and* arcs) that start from different *premise* nodes and converge at a single *consequence* node, and is identified by a curved segment that joins the arcs from premises, meaning that all premises are necessary to reach the consequence; the *disjunction* is a relationship between multiple nodes, represented by arcs (*or* arcs) that start from different *premise* nodes and converge at a single *consequence* node, and is not specifically identified, meaning that a single premises is sufficient to reach the consequence. This type of graph is not supported by algorithms like Dijkstra, that would choose the single lower weight arc out of a whole *and* group. To address the problem, the *hypergraph* structure was adopted.

The peculiarity of hypergraphs is that an hyperedge can connect arbitrary subsets of vertices, and not only pairs. This modeling is more suited to our case study, since as a learning activity can entail more different skills required (source nodes) and more acquired skills (destination nodes). The AO\* algorithm was implemented to find the minimum cost subgraph in the new adopted structure. This algorithm searches the optimal solution for the minimum path in an and-or graph based on a heuristic function, which represents an estimate of the overall cost of the solutions, and on a cost function, which represents the actual cost of the path that starts from a node and arrives to all the solutions. As a heuristic function, in our case, the minimum weight of the incoming edges to a node is used. Our oriented hypergraph is composed of an hyperedge for each learning activity  $la$ , that stores source nodes (the set  $la.P$ ) and destination nodes (the set  $la.A$ ), and is labelled by the cost of the learning activity, and by a node for each skill that stores the entering (hyper)edges (learning activities that have those skills in  $la.A$ ) and out (hyper)edges (learning activities that have those skills in  $la.P$ ). As in general AO\* solution, the first step is a top-up down, starting from the root (the new skill) and reaching a leaf (a skill already in  $SK(l)$ ) by traversing the nodes found in the partial solution; the second step is bottom-up, and

propagates towards the root the information from the leaf, by expanding unvisited nodes and updating antecedent nodes cost based on subsequent nodes cost, and marking the best path.

The AO\* algorithm is integrated with computations useful for the solution of our problem. In particular, during the search of the subgraph of minimal cost and for the estimation of the ZPD, *AvgCertainty*, *AvgEffort* and *DThreshold* values are computed, as discussed in the following section

## VI. ZPD ESTIMATION AND VISUALIZATION

Given a course *LP*, its knowledge domain is  $KD(LP) = LP.A \cup LP.P$ . In particular,  $KD(LP) \setminus s\text{-proj}(SK(l))$  is the set of all skills in the course knowledge domain, that are not yet in  $SK(l)$ . According to the above definitions, such skills would belong neither in  $APS(l)$  nor in  $ZPD(l)$ . Given one such skill, and given its distance from  $SK(l)$  and the support set of the path determining such distance as defined above, it is reasonable to assume that the more the skills in the support set and the higher the certainty associated to them, the easier it is possible to expect the learner to reach *s*. Likewise, certainty in the support set can be used to estimate how far from the  $SK(l)$  the student can go trying to acquire new skills, and yet still consider these skills in the  $ZPD(l)$ . The concrete possibility for a student to achieve a certain skill does not only depend on the required effort, but also on the certainty of the elements supporting such achievement. To clarify this point, let us suppose  $D(s, s\text{-proj}(SK(l), LP)) \geq D(s', s\text{-proj}(SK(l), LP))$ : if the overall certainty of  $Support(s, SK(l), LP)$  is higher than  $Support(s', SK(l), LP)$ , we can conclude that *s* might be reachable while *s'* might not, in spite of a closer distance. Therefore, the estimation of the inclusion of a skill *s* in the ZPD of a student *l* requires a more detailed reasoning. Once we succeed in finding an effective/suitable definition for the two following measures:

$$A1 = AvgEffort(G^*, Support(s, SK(l), LP))$$

as an estimate of the “average” effort required by each activity in  $G^*$ , and

$$A2 = AvgCertainty(Support(s, SK(l), LP))$$

as an estimate of the “average” certainty of the skills in the support set for  $G^*$ , we will be able to use them to define a suitable *daring threshold*, i.e., the maximum distance from a specific learner’s state of knowledge  $SK(l)$  of a certain skill *s*, below which it is acceptable that *s* is in  $ZPD(l)$ ; this threshold can reasonably be computed as:

$$DThreshold(s, SK(l)) = (A2/A1) \cdot EFF(R) \cdot dF$$

where the ratio represents in some sense the amount of certainty per unit of effort,  $EFF(R)$  relates the threshold with an estimation of the average effort in the repository, and  $dF$  is the *daring factor*, an integer regulating how “far from the present skills (the present state of knowledge) it is pedagogically acceptable/sustainable to go and place the ZPD boundaries. In practice,  $dF$  is used both to “normalize” the quotient factor in the above expression (so to be comparable with distances) and to provide the system (and ultimately the teacher) with a way to

configure the maximal distance of the ZPD boundaries from one’s possessed skills. Then we can then define the ZPD as the set of skills, in the portion of knowledge domain entailed by the learning path/course tackled by the learner, whose distance from the skills in the learner’s state of knowledge is within the above stated threshold

$$ZPD(l) = \{s \in KD(LP) \setminus APS(l) \mid D(s, s\text{-proj}(SK(l), LP)) \leq DThreshold(s, SK(l))\}$$

It is worth noticing that this definition of the ZPD is highly dynamic, personalized, and adaptive, since its radius depends on the course activities, on the student, and on the student’s state of knowledge, and furthermore entails a boundary determination that depends on the specific skills at hand.

Thanks to the above dynamic estimation of ZPD, given an attended course, the student can have a graphical overall vision of skills that can be reached, of those already present in the cognitive state, and of those which are still too far away. This supports the metacognitive activity of self-evaluation. In particular, the course is presented as a graph, with nodes being skills (green, yellow or red according to their inclusion in SK, ZPD or UPS) and edges labelled with activities entailing such skills as prerequisites or acquirement. Further details can be asked on demand, on both skills (e.g., present distance from SK) and activities (e.g., estimated effort). Figure 2 shows an example of graph with skill details on the right frame.

The work in [4] presented a first proposal for the computation of the two core values  $A1$  and  $A2$  above. Much of the rest of this work deals with refining that proposal.

## VII. A NEW DEFINITION OF *AVGEFFORT()* AND *AVGCERTAINTY()*

The function *AverageCertainty()* is an important ingredient for the ZPD computation. It estimates, for each new skill possibly joining in ZPD, an overall estimated value for the certainty of a support group of prerequisite skills. The main principle is that, in such a computation, a bare average is not the most reliable value. Rather, each possessed skill  $s_i$  should intervene with a specific weight  $w_i$  computed for its certainty  $c_i$ . Such weight should depend on aspects that make the possessed skill more or less relevant in the current cognitive state of the learner with the aim to reach the new skill. Different skills might contribute differently to this aim: those with different certainty can be expected to contribute differently, and different skills presenting the same certainty might contribute differently too, depending on other characteristics connected to pure certainty.

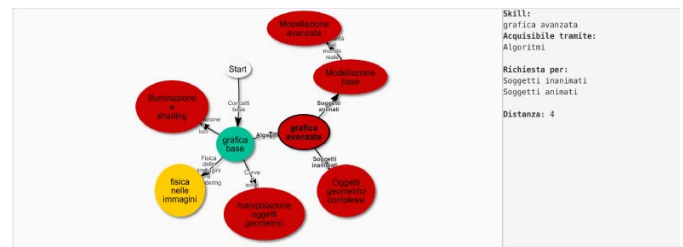


Figure 2. Example of a graph with skills coloured according to their *reachability*. The right frame shows further details for a selected skill.

In particular, we do not consider a “semantic” viewpoint, but rather the additional characteristics listed below, that take into account the consolidation of a skill in time and the ways that certainty has been reached during the learning work. For sake of the reader, we report them as introduced in [4] and the way they are used there to compute  $w_i$ :

- **age( $s_i$ ) Age of the skill:** the time elapsed from the moment the learner first acquired it;
- **age(cert( $s_i$ )) Age of the present certainty value:** the time elapsed from the moment the learner got the present value of certainty for a skill;
- **age( $s_i$ ) - age(cert( $s_i$ )) Time from first acquisition to present certainty level:** the time needed to reach the present level of skill (progress speed);
- **ntests/npostests ratio The rate between the number of tests and the number of positive increments of certainty:** how many times the learner achieved a positive result on a test related to the skill, compared with the total number of tests related to the skill.

Given a learner  $l$  and a skill  $s_i$  from  $SK(l)$  with certainty  $c_i$ , the weight for  $\langle s_i, c_i \rangle$  as computed in [4] is

$$W_i = \frac{\text{age}(s_i) * \text{age}(\text{cert}(s_i))}{(\text{age}(s_i) - \text{age}(\text{cert}(s_i))) * (\text{ntests} / \text{npostests})}$$

where being on the numerator or denominator depends on the deemed either direct or indirect proportionality to the proposed weight (for detailed motivations, see [4]). Notice that the ratio between the number of verifications operated on the skill –  $ntests$  – and the number of them that were successful –  $npostests$  is equal to 1 for the straight and all-successful path (no effect on the weight), and progressively higher when the percentage of unsuccessful verifications increases (weakening the weight accordingly more).

The value computed for the support set of a skill outside  $SK$  is therefore:

$$A1 = \text{AvgCertainty}(\text{Support}(s, SK(l), LP)) = \frac{(\sum_{\langle s_i, c_i \rangle \in \text{Support}(s, SK(l), LP)} W_i \cdot c_i) / \text{Card}(\text{Support}(s, SK(l), LP))$$

Average certainty can be obtained through a backward computation, from a skill to be possibly included in ZPD, towards its support set. On the contrary, the computation of an average effort can be obtained by a forward computation from a support set towards any reachable skill to be possibly included in the ZPD. We can assume that, for each  $la$  in  $G^*$ , a subset of skills in both  $la.P$  and  $la.A$  are already in  $SK(l)$  and have reached some certainty level. This is accounted for while computing the  $\text{AvgEffort}$  required by the student along a learning path. In other words, each activity has an effort value included by the instructional designer in its definition, but this value might be actually different in the practical experience of each single student  $l$ . This depends on the possible presence in  $SK(l)$  of some of the skills included in the acquired  $la.A$ , and on the reached level of certainty of the activity’s prerequisite  $la.P$ .  $\text{AvgEffort}$  is transformed in a weighted sum. The computation of the weight  $w(la)$  for  $la$ , takes into account

skills in both  $la.P$  and  $la.A$  which are already in  $SK(l)$  and their certainty. Since part of the path is outside  $SK$ , it happens that some  $P$  sets on the way are completely out of current  $SK$  and are filled while proceeding in the learning path. Skills already possessed, both in prerequisites and to-acquire sets, contribute to decrease the effort actually experience by the student. But it is appropriate to consider skills in  $la.P$  and  $la.A$  in a different way. Since the expected level of certainty for a newly acquired skill is  $C_{ENTRY}$ , any higher level of certainty already achieved can facilitate the task. On the contrary, if for some prerequisite skill the level is under  $C_{PROMOTE}$ , towards  $C_{DEMOTED}$ , this will possibly increase the effort value. We define  $w(la)$  according to the following:

$$w(la) = \frac{wa(la.A) + wp(la.P)}{2}$$

with

$$wa(la.A) = \text{Card}(la.A) \cdot c_{\text{entry}} / \sum_{s \text{ in } la.A} f(s, SK(l))$$

$$wp(la.P) = \text{Card}(la.P) \cdot c_{\text{promote}} / \sum_{s \text{ in } la.P} f(s, SK(l))$$

where

$$f(s, SK(l)) = \begin{cases} c & \text{if } \langle s, c \rangle \in SK(l) \\ C_{ENTRY} & \text{otherwise} \end{cases}$$

notice that  $wa(la.A)$  will assume value 1 when none of the skills to acquire is already possessed, otherwise will be less than 1 so decreasing the value of the effort for that activity in the sum; as for  $wp(la.P)$ , it will assume its minimum value 1 when all the skills are already possessed with certainty  $C_{PROMOTE}$ , otherwise it will be greater than one and therefore increase the effort, in a way proportional to the distance from full certainty. In most cases, no skill to acquire is already possessed ( $wa(la.A) = 1$ ) and prerequisite skills are expected to be possessed with certainty  $C_{PROMOTE}$  ( $wp(la.P) = 1$ ) and therefore  $w(la) = 1$  and the value of the effort for the student is not affected. For more discussion see [4]. The weighted sum  $\text{AvgEffort}$  is:

$$A2 = \text{AvgEffort}(G^*, \text{Support}(s, SK(l), LP)) = \frac{\sum_{la \in G^*} w(la) \cdot la.\text{effort} / \text{Card}(G^*)$$

While no modification is proposed for the definition of  $\text{AvgEffort}()$ , we will compare the previous computation of  $\text{AvgCertainty}()$  with a new one.

The first consideration is that the age of a skill lends itself to contrasting considerations. From one side, as argued in the previous work, an “elder” skill should be more consolidated/firm, and then achieve a higher weight. On the other hand, if not suitably exercised, the same skills might weaken through time, taking to opposite considerations. Therefore, we decided to test the deletion of  $\text{age}(s_i)$  from the weight computation, and to substitute it with the number of test successfully carried out related to the skill. While this value is already present in the ratio  $ntests/npostests$ , we tested the result of giving it an autonomous role in the weighting

formula. In other words, we introduced a direct proportionality with the value  $npostests^2/ntests$ . A further consideration regards  $age(cert(s_i))$ . It is certainly true that if its certainty has been updated recently, then that skill has been used by the student on a recent date. Consequently, a high  $age(cert(s_i))$  should decrease the weight, while a low one should increase it. This factor must still be modulated by the value  $npostests/ntests$ , which is the actual information about the overall trend of student's performance relative to  $s_i$ , and somehow accounts for possible decreases in its certainty value. Therefore, we have an inverse proportionality with  $age(cert(s_i)) \cdot ntests/npostests$ . Summarizing, the new weight for skill  $s_i$  in the computation of  $AverageCerativity()$  is

$$w_i = \frac{npostests^3}{age(cert(s_i)) \cdot ntests^2}$$

In order to carry out an in-vitro comparison of the two weighting formulas, we applied them to the situation depicted in Table 1, listing the learning activities in some hypothetical course, with related skills (Prerequisite and Acquired) and estimated effort in the activity description. To simplify the example, we assume that  $la.P$  and  $la.A$  are singletons.

TABLE 1 AN EXAMPLE COURSE CONFIGURATION

Learning act.	la.P	la.A	effort
cb	/	gb	5
il	gb	is	10
fir	gb	fni	2
cs	gb	mog	7
a	gb	ga	4
si	ga	ogc	6
sa	ga	mb	9
smr	mb	ma	8

Table 2 and Table 3 show the ZPD of a beginner student with  $SK(l) = \{ < gb, 5 > \}$ , for different values of the elements appearing in the two weighting formulas. The conceptual difference between the two formulas is particularly evident if we consider the graphical representation of the maximum resulting ZPDs as they would be shown to the student (Figures 3 and 4 respectively).

TABLE 2 RESULT ZPD WITH THE FIRST FORMULA FOR  $AverageCerativity()$

$age(s_i)$	$age(cert(s_i))$	$npostests/ntests$	$w_i$	ZPD(l)
1 week	1 week	1	0.1	{ is, fni }
1 week	1 week	1/2	0.05	{ fni }
1 week	1 week	1/4	0.025	∅
1 week	1 week	3/4	0.075	{ fni }
3 weeks	1 week	1	0.3	{ is, fni, ga }
3 weeks	1 week	1/2	0.15	{ is, fni }
3 weeks	1 week	1/4	0.075	{ fni }
3 weeks	1 week	3/4	0.225	{ is, fni, ga }
3 weeks	3 week	1	0.9	{ is, fni, mog, ga, ogc, mb }
3 weeks	3 week	1/2	0.45	{ is, fni, ga }
3 weeks	3 week	1/4	0.225	{ is, fni, ga }
3 weeks	3 week	3/4	0.675	{ is, fni, mog, ga, ogc }

TABLE 3 RESULT ZPD WITH THE SECOND FORMULA FOR  $AverageCerativity()$

$age(cert(s_i))$	$npostests/ntests$	$w_i$	ZPD(l)
1 week	1	0.1	{ is, fni }
1 week	1/2	0.025	∅
1 week	1/4	0.006	∅
1 week	3/4	0.168	{ is, fni, ga }
2 week	1	0.05	{ fni }
2 week	1/2	0.012	∅
2 week	1/4	0.005	∅
2 week	3/4	0.084	{ fni }
3 week	1	0.033	{ ∅ }
3 week	1/2	0.008	∅
3 week	1/4	0.002	∅
3 week	3/4	0.056	{ fni }

Notice that in both tables the obtained results are consistent with the direct and indirect proportionalities devised for the different elements. On the other hand, in both cases it is possible to observe that, in general, the student can diverge from a strictly linear path, by choosing at any moment among different activities, according to the preference, to the current resources and, perhaps, to the available time (e.g., it is possible to choose the activity currently requiring a lower effort to be able to fulfil more commitments). In comparison with the present deployment of learning material even on advanced platforms, it is clear how student's personal responsibility and ability to self-regulate the content and pace of learning are supported.

Actually, the second formula seems much more restrictive, but, also considering the course structure in Table 1, it seems that the result in Figure 4 (second formula) is more reasonable, due to the fact that "double jumps" are not allowed, i.e., is a prerequisite is missing it is suggested to the student not to tackle a certain activity. However, it is also possible to observe that an activity for which the prerequisite has been achieved is not included in the ZPD. Moreover, it is possible to observe an anomaly when the larger ZPD set is identified with a lower  $npostests/ntests$  ratio.

The above results and observations suggest the need for further investigations and, most of all, for a real field experimentation.

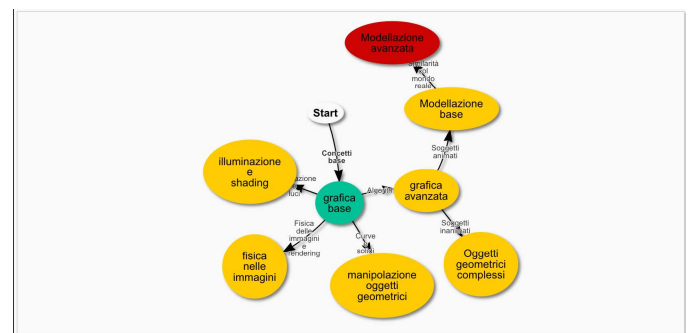


Figure 3 Maximum achievable ZPD with the first weighting formula for  $AverageCerativity()$

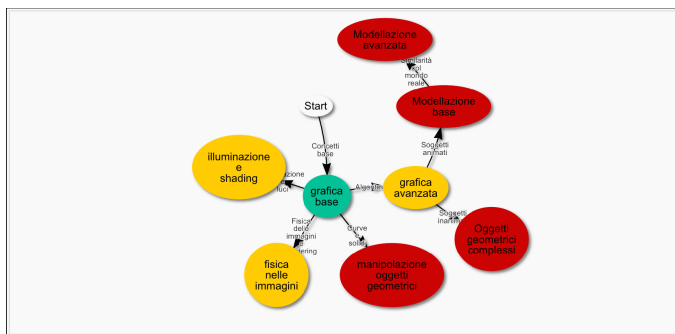


Figure 4 Maximum achievable ZPD with the second weighting formula for *AverageCeratinty()*

### VIII. CONCLUSIONS

The theory by Vygotskij is inherently a social one, therefore it is mandatory to extend the computation of ZPD to a group dimension. A discussion of the possible choices is presented in [17]. In that work, a compromise is proposed between a too narrow ZPD, resulting too boring for smart students, and a too large one, penalizing weak ones. Ongoing investigations are being carried out, that cannot be reported here for sake of space. However, the new obtained results are very promising in a group perspective too. The considerations from which the present work stems are twofold. First, in present distance learning strategies the only parameter that is taken into account to evaluate the “reachability” of a learning activity is a standard (average) effort hypothesized by the activity creator. Notwithstanding the educational experience underlying such hypothesis, the real effort experienced by a student heavily relies on her/his cognitive state at the moment the activity is actually tackled. As a consequence, a lighter or heavier effort can be required from different students, depending on the already possessed abilities that were expected to be acquired with the activity, or on the prerequisite ones that are still lacking. This is the core consideration that drives the present estimation of a pedagogically relevant set of activities that can be fruitfully tackled by each student. These activities compose what we deem an operational equivalent of the theoretical concept of ZPD by Vygotskij, though bound to an individual dimension. A second consideration regards the possibility provided by the devised tools to enforce the meta-cognitive abilities of self-assessment and self-regulation. Providing the student with a reliable set of activities that can be fruitfully tackled, can both leave more freedom, therefore increasing student’s motivation in tackling the preferred activity at any moment, and eliminate as much as possible the risk of being frustrated by a too difficult task. At the best of our knowledge, this kind of transparent guidance has never been implemented in any distance learning system.

### REFERENCES

- [1] L. S. Vygotskij, “The development of higher forms of attention in childhood,” in J. V. Wertsch (Ed.), *The concept of activity in Soviet psychology*. Armonk, N.Y.: Sharpe, 1981.
- [2] M. De Marsico, A. Sterbini, and M. Temperini, “The definition of a tunneling strategy between adaptive learning and reputation-based group activities,” *Proc. of 11th IEEE International Conference on Advanced Learning Technologies (ICALT 2011)*, pp. 498-500, July 2011.
- [3] M. De Marsico, A. Sterbini, and M. Temperini, “A strategy to join adaptive and reputation-based social-collaborative e-learning, through the zone of proximal development,” *International Journal of Distance Education Technologies (IJDET)*, 11(3), pp. 12-31, 2013.
- [4] M. De Marsico and M. Temperini, “Average effort and average mastery in the identification of the Zone of Proximal Development,” *6th International Workshop on Social and Personal Computing for Web-Supported Learning Communities (SPeL 2013)* Sinaia, Romania 11-13 October 2013, pp. 651-656.
- [5] M. De Marsico, A. Sterbini, and M. Temperini, “Assessing Group Composition in e-learning According to Vygotskij’s Zone of Proximal Development,” in: C. Stephanidis and M. Antona (Eds.): *UAHCI/HCI 2014*, 27 June 2014, Heraklion, Crete, Greece. Universal Access in Human-Computer Interaction. Universal Access to Information and Knowledge, Part II, LNCS 8514, pp. 277–288, 2014.
- [6] P. Brusilovsky and J. Vassileva, “Course sequencing techniques for large-scale web-based education,” *International Journal of Continuing Engineering Education and Life-long Learning*, 13, pp. 75–94, 2003.
- [7] P. Brusilovsky and E. Millan, “User models for adaptive hypermedia and adaptive educational systems,” in P. Brusilovsky, A. Kobsa, and W. Nejdl (Eds.), *The Adaptive Web: Methods and Strategies of Web Personalization*, LNCS 4321, pp. 3–53, 2007. Springer-Verlag, Berlin Heidelberg New York.
- [8] S. Graf and Kinshuk, “Providing adaptive courses in learning management systems with respect to learning styles,” in *World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education (eLearn 2007)*, pages 2576–2583, 2007.
- [9] C. Limongelli, F. Sciarrone, M. Temperini, and G. Vaste, “Adaptive Learning with the LS-Plan System: a Field Evaluation,” *IEEE Transactions on Learning Technologies*, vol. 2, no. 3, pp. 203-215, 2009.
- [10] P. De Bra, D. Smits, and N. Stash, “Creating and delivering adaptive courses with AHA!” in (Nejdl, W. and Tochtermann, K. Eds. ) *Innovative Approaches for Learning and Knowledge Sharing - First European Conference on Technology Enhanced Learning, EC-TEL*, pages 21–33, LNCS 4227, 2006.
- [11] K. Silius and A. Tervakari, “The usefulness of web-based learning environments,” *Proceeding of the International Conference on Network Universities and e-learning*, 2003.
- [12] K. Kreijns, P.A. Kirschner, W. Jochems, „Identifying the pitfalls for social interaction in computer supported collaborative learning environments: a review of the research,“ *Computers in Human Behavior*, 19, pp. 335-353, 2003
- [13] Y. Cheng, and H. Ku, “An investigation of the effects of reciprocal peer tutoring,” *Computers in Human Behavior*, 25(1), pp. 40-49, 2009.
- [14] A. Leblanc, M.-H. Abel, „E-MEMORae2.0: an e-learning environment as learners community support,“ *Int. Jou. of Computer Sciences & Applications*, vol. 5 no.1, pp. 108-123, 2008.
- [15] L.S. Vygotskij, *Mind in society: the development of higher psychological processes*. (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds. and Trans.) Cambridge MA: Harvard University Press, 1978. (Original material published between 1930 and 1935).
- [16] B.S. Bloom (Ed), *Taxonomy of Educational Objectives*, David McKay Company Inc., New York, 1964.
- [17] M. De Marsico, A. Sterbini, M. Temperini. *Assessing Group Composition in e-learning According to Vygotskij’s Zone of Proximal Development*. In: C. Stephanidis and M. Antona (Eds.): *UAHCI/HCI 2014*, 27 June 2014, Heraklion, Crete, Greece. Universal Access in Human-Computer Interaction. Universal Access to Information and Knowledge, Part II, LNCS 8514, pp. 277–288, 2014.