Resource-Rich versus Resource-Poor Assessment in Introductory Computer Science and its Implications on Models of Cognition: An in-Class Experimental Study

Tobias Halbherr[1, 2] (tobias.halbherr@gess.ethz.ch), Hermann Lehner[3] (hermann.lehner@inf.ethz.ch), Manu Kapur[1] (manukapur@ethz.ch)

[1]ETH Zurich; Department of Humanities, Social and Political Sciences; Institute of Learning Sciences and Higher Education

[2]ETH Zurich; Educational Development and Technology

[3]ETH Zurich; Department of Computer Science

Abstract

Outside university, students encounter disciplinary practices mediated by technological resources. In this sense, the real world is decidedly resource-rich. In contrast, most educational assessments remain decidedly resource-poor. Situated versus mindbased perspectives of cognition fundamentally differ in the role they ascribe to such resources in cognition and learning. To mindbased perspectives, they are a source of input, to situated perspectives they are constitutive to cognition itself. We assessed the validity of resource-rich versus resource-poor assessments of learning outcomes from resource-rich versus resource-poor learning activities. The study implemented an in-class 2x2 between-subjects experimental design in an introductory programming course with 192 first semester BSc engineering students. Both types of assessment were sensitive to differences in learning outcomes, indicating validity for both. Results indicate resource-rich assessments may be more ecologically valid, while – intriguingly – the resource-poor assessments were more sensitive to transfer of learning. Furthermore, the resource-rich learning activities better facilitated learning for transfer.

Keywords: assessment; examinations; resource-rich assessment; resource-affordances; higher education; learning science; computer science education; e-assessment; educational technology; situated cognition

Introduction

Examinations in (higher) education usually remain restricted to pen and an empty piece of paper – or in their computer-based counterpart, keyboard, mouse, and a standardized e-assessment environment. What examinations typically lack – indeed prohibit – is access to any additional resources. In this sense, conventional examinations are *resource-poor*. In contrast, upon leaving university students will usually have access to a wide array of resources, such as specialist tools, easy access to information, and support from networks of experts and peers. In this sense, most professional practices outside the classroom are decidedly *resource-rich*. However, if the practices in the real world – for which we ultimately learn – are resource-rich, how can we justify a resource-poor examination practice? Conversely, how could we demonstrate the need for examinations to become resource-rich? In this study, we render first empirical evidence unto this question for the case of *tools* as resource. Specifically, we are interested in the question whether the availability or absence of disciplinary technological tools in an assessment environment has implications on the *validity* of the corresponding assessments of learning.

Our research question lies at the intersection of three larger topics which to date have rarely been linked. First, the above-mentioned discrepancy between (increasingly) resource-rich disciplinary practices versus resource-poor conventional examination practice and its implication on validity. Second, the resurgent epistemic debates on appropriate perspectives of cognition and learning. Third, advancements in educational technology, which enable novel learning and assessment environments. We will briefly elaborate on each.

Resource-Rich Assessment

When asked to formulate intended learning outcomes, lecturers typically emphasize outcomes associated with deep learning, transferrable skills, and rich conceptual understanding. Conventional examinations in contrast, are frequently associated with surface learning, cramming, factual recall, poor retention, and an inability to transfer (Biggs, 2014; Keehner, Gorin, Feng, & Katz, 2017). In other words, there seems to be a problem with the validity of conventional examinations: Lecturers intend to assess outcomes associated with deep learning, but in effect, students may achieve success through surface learning. To make matters worse, examinations strongly motivate student learning and when examinations reward surface learning, they also encourage surface learning. Assessment drives learning (Baird, Andrich, Hopfenbeck, & Stobart, 2017) and poor assessment drives poor learning.

Alternative Assessment (Sambell, McDowell, & Brown, 1997), Authentic Assessment (Gulikers, Bastiaens, & Kirschner, 2004), Assessment for Learning (Baird et al., 2017), or Performance Assessment (Moss, 1992) all share with our proposition for resource-rich assessment a concern for the above-mentioned issues with assessment validity and/or assessment driven learning. However, none of these approaches foreground the access to relevant disciplinary resources (tools, information, and/or social interactions). We propose that the absence of relevant disciplinary resources in assessment contexts may be a crucial mediator of longstanding issues with both assessment validity and assessment driven learning. We propose three principal reasons for this. First, technological resources mediate and pervade an ever-increasing number of disciplinary practices: Computer scientists develop code in integrated development environments (IDEs), psychologists do statistics in R,

engineers design machine parts in CAD software, medical practitioners treat and diagnose patients with the aid of clinical decision support software, and virtually everyone writes texts with word processors. Second, learning sciences research indicates that successful, transferrable learning is associated with learners' active engagement with appropriate tools and learning resources (e.g. Danish & Gresalfi, 2018; Schwartz & Martin, 2004; (Hmelo-Silver, Kapur, & Hamstra, 2018). If successful learning is resource-mediated, then the resource-poorness of conventional examinations may explain issues with assessment driven learning: Resource-poor assessment may drive resource-poor learning. Third, cognition itself may be substantially resource-mediated.

Cognition

Established examination practice and its frameworks have been criticized for paying too little attention to the cognitive models in which they are grounded (Baird et al., 2017), and/or for being based on impoverished, outdated, or unsuitable models of cognition (Pellegrino, 2002; Sawyer, 2014). For the purpose of this study, we compare a rigid interpretation of two contrasting perspectives of cognition: Cognition as *mindbased* processing versus cognition as *situated* action. The mindbased perspective corresponds to cognitivist (sic), computational, representational, information-processing, connectionist, or constructivist models of cognition and learning (Abrahamsen & Bechtel, 2012; Shapiro, 2011). The mind is the manifest locus of cognition and learning, and mediator of the relationship between stimuli and response. Fundamental to this perspective is the 'opening of the black box' by modelling processes and states within the mind. Two simultaneous and intertwined streams of processing in the mind/brain together constitute cognition: Directed feedforward sensory-to-motor, stimulus-response, input-output streams of processing in combination with recursive feedback loops within the mind/brain itself. The contrasting situated model on the other hand, does not separate cognitive processing ('mind') from action ('response') or social and physical task contexts ('stimuli'). Instead, it regards the dynamical interaction between the cognitive agent and those elements of the environment with which he/she situationally interacts as conjointly constitutive of cognition and learning: Cognition as *situated action* or as *emergent* upon loosely coupled processes in the agent-environment complex system (Clark, 2012; Danish & Gresalfi, 2018; Hutchins, 1995). Actions of the cognitive agent effect changes in the environment, which in turn feed back to the cognitive agent in the form of new/altered stimuli. The directed stimulus-response flow of processing of the mindbound perspective is closed into a single complex system of dynamical feedback loops, from the cognitive agent through the environment back unto the cognitive agent him/herself. Examples of situated perspectives include embodied, extended, and distributed cognition, sociocultural theory, or social constructivism. While these situated perspectives have led to a rich body of research on learning and effective learning interventions, there is a lack of corresponding research in assessment and educational measurement (Mislevy, 2018 is one exception).

Gibson (1977) introduced the term 'affordance' for "whatever it is about the environment that contributes to the kind of interaction that occurs [with the cognitive agent]". Accordingly, we define the term *resource-affordance* for 'whatever it is about the disciplinary (technological) resources with which the cognitive agent interacts, that contributes to the kind of disciplinary practice that occurs'. Resource-affordances constitute the loose coupling of processes between the cognitive agent and the task environment. They are fundamental to the situated perspective. Much like a skier's body is inseparably connected with his boots and skis in the practice of skiing – effectively forming a single functional unit – so too do a programmer's mind and a computer-based programming environment interact in an inseparable manner in the practice of programming. Just as attempting to assess someone's skills in skiing while denying them skis would be rather absurd, so it is absurd that we routinely assess students' competency in computer science while denying them access to computers. It follows that the valid assessment of competency in disciplinary practices directly depends on adequate access to practice-relevant resource-affordances in the assessment task environment. Hence, the situated perspective demands an examination practice that is equally resource-rich (RR) as are the disciplinary practices in which we intend to assess competency. In the mindbased perspective on the other hand, there is no need to model resource affordances because cognition is fully contained in the mind/brain. Writing a recursive algorithm on paper or in a programming environment are not fundamentally different cognitive tasks, but fundamentally similar. Resources do not contribute anything substantial to cognition or its assessment. On the contrary, they are a potential source of construct irrelevant variance. Hence, the mindbased perspective of cognition favors a resource-poor (Rp) assessment practice. Indeed, we argue that a main reason for conventional examination practices being resource-poor likely lies in the fact that most students, teachers, educators, and assessment specialists share a deeply mindbased conception of cognition.

Educational Technology

Over the past years, educational technology and corresponding e-learning practices, including computer-based assessments and examinations, have become increasingly widespread in higher education (Bennett, 2015; Crisp, Guàrdia, & Hillier, 2016; Halbherr, Reuter, Schneider, Schlienger, & Piendl, 2014). Computer-based assessment services frequently prioritize efficiency by focusing on auto-correction, computer-adaptive testing, or remote proctoring – all largely within a conventional Rp paradigm. However, there also exists a competing trend, emphasizing the potential for improvements in examination quality by enabling examination task environments that are more authentic, competence-oriented, aligned with corresponding practice – and/or RR (Crisp et al., 2016; Halbherr, Dittmann-

Domenichini, Piendl, & Schlienger, 2016). One example of such a learning and assessment environment is Code Expert.

Code Expert

Code Expert (Lehner, Avanthay, & Sichau, 2018) is a browser-based integrated development environment (IDE) and online learning environment developed at ETH Zurich. Code Expert facilitates open programming assignments for in-class or take-home exercises, as well as supervised examinations. Code Expert includes an auto-grader, which provides automatic and immediate feedback to students by compiling, running, and testing submitted code against predefined test cases. Furthermore, tutors can annotate or apply direct changes to students' attempts in order to provide additional, more personalized feedback. The Code Expert interface is illustrated in Figure 1. It consists of a file system pane, a code editor window, a terminal and output window for compiling and running the code, and a tutorial pane for instructions, task descriptions, and learning materials.

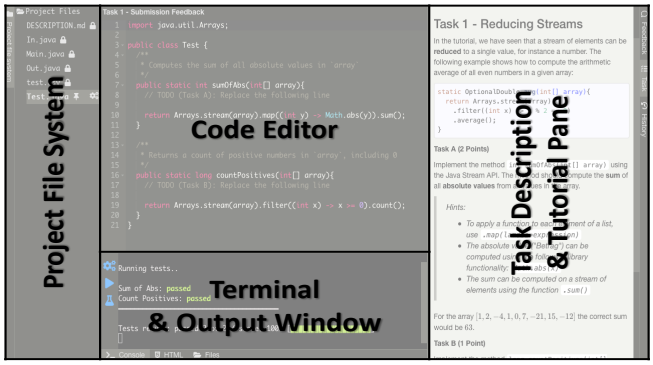

Figure 1: Schematic overview of Code Expert GUI.

Method

The study was conducted as part of an introductory course in programming for non-CS students at first year BSc level. The course focuses on imperative and object oriented programming paradigms, as well as problem solving, and uses Java as programming language and Code Expert as learning environment.

In the study, we investigated how the presence or absence of resource-affordances of the Code Expert environment affected student learning on the one hand, and the assessment of corresponding learning outcomes on the other. Slightly different than usual, the main focus of this study is not on the learning activity, but on the assessments, more precisely: The *validity* of the assessments. Specifically, we are interested whether and to what extent RR versus Rp assessments are sensitive to differences in learning outcomes as induced by RR versus Rp learning activities.

Validity

Validity is "the degree to which a test or examination measures what it purports to measure" (Ruch, 1924). It is a unitary construct (Messick, 1989). It is an ontological and/or epistemic construct, rather than a statistical or psychometric one (Kane, 2006). This holds particularly true in the context of this study, since we do not have any impartial source of base truth against which we could validate the RR versus Rp assessments. Instead, validity has to be determined through an appropriate validity argument. Borsboom, Mellenbergh, & van Heerden (2004) propose the following operational definition of validity: "A test is valid for measuring an attribute if and only if (a) the attribute exists and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure".

Operationalizing Validity

We apply the earlier propositions – resource mediation facilitates learning and cognition as a resource-mediated construct – to the above operational definition of validity. We operationalize variations in the measurement attribute – student learning – by letting one group of students learn with access to relevant resource-affordances, the RR learning condition (LC), while the other group is denied access, the Rp LC. Everything else is kept strictly identical across the two LCs. Subsequently, we assess half of students of each LC in a RR assessment with access to resource-affordances, the RR assessment condition (AC), or in a Rp assessment without access, the Rp AC. Again, everything else is kept strictly identical across the two ACs. We then evaluate whether the RR and/or the Rp assessment are able to differentiate between students from the RR LC versus the Rp LC. If they do, then this is evidence in favor of the assessment's validity, and evidence against, if not. This results in a 2x2 between-subjects experimental design. The RR LC versus Rp LC and the RR AC versus Rp AC constitute the independent variables, and assessment performances – to be more precise, the performance differences between the students in the RR LC and the Rp LC as measured either in the RR AC or the Rp AC – constitute the independent variables.

Operationalizing the Resource-Affordances

We identify the compiler as the key resource in the Code Expert environment. The compiler is both essential for generating the product of the practice – running code – as well as for sustaining the practices and processes required for achieving that goal. We thus operationalize the RR experimental conditions with a Code Expert environment with a fully functional compiler. The Rp conditions we operationalize with the *identical* Code Expert environment save for a deactivated compiler. This leads to the disappearance of the following resource-affordances: Students cannot compile or run code, correspondingly cannot receive any messages in the console from either the compiler or their compiled code, cannot run their code against test cases in the auto-grader, and there is no syntax highlighting of their code in the code editor. In the Rp condition, the students are essentially working with a 'naked text editor version' of Code Expert, while in the RR condition they have access to the fully functional Code Expert IDE. Across all

experimental conditions, students were instructed not to access any other resources (e.g. lecture notes, Google, StackOverflow, other Code Expert exercises) than those available through the study tasks in Code Expert.

Learning and Assessment Tasks

In the study's learning activity, we introduced a new paradigm: Functional programming. Java implements functional programming with the Stream API. The learning activity consisted of an interactive self-study tutorial. Key concepts of functional programming were introduced and consolidated in five consecutive tasks using hands-on exercises with the example of manipulating data-streams of numbers. Students received the canonical solution to each tutorial task at the start of the subsequent tutorial step. This ensures that also the students in the Rp LC received adequate feedback on the correctness of their solutions.

The subsequent assessment consisted of three tasks. In Task1 students had to perform identical manipulations of data-streams of numbers as in task 4 of the tutorial. Task1 operationalizes the direct replication of learning. Task2 introduced a novel and more complex problem that can be solved elegantly using the new paradigm. Task2 operationalizes transfer of learning. In Task3, students had to manipulate streams of Java objects instead of streams of numbers after reading a brief introduction to a number of new concepts and operations for manipulating objects in a functional manner. Task3 operationalizes transfer of learning with the aid of a learning resource, i.e. students' preparation for future learning (Schwartz & Martin, 2004). All three assessment tasks were scored manually by the course assistants. For each task 0, 1, 2, or 3 points were awarded according to task-specific rubrics. Small syntax errors, such as missing or unmatched brackets or slightly incorrect syntax in lambda expressions, were ignored. The manual scoring procedure was identical for both the RR AC and the Rp AC. To ensure a high correspondence with actual educational practice 'in the wild', both the learning activity and the assessment, all tasks contained therein, the scoring rubric, and the scoring procedure were all designed and performed entirely by the course lecturer and the course assistants, with no or only minimal intervention from the lead investigator.

Hypotheses from the Cognitive Perspectives

Let us now revisit the situated versus mindbased perspectives of cognition. What kind of results would each perspective predict for this experiment? To the situated perspective, the loose coupling of processes through resource affordances remains intact in the RR LC and the RR AC, while in the Rp conditions this coupling is severed. Hence, the RR and the Rp experimental conditions correspond to qualitatively fundamentally different kinds of cognitive processes – both regarding what is learned in the LCs, as well as regarding what is assessed in the ACs. Since programming is a resource-mediated practice, the situated perspective would predict larger learning gains in the RR LC, to which the RR AC is sensitive, but not the Rp AC (or only to a lesser extent). Furthermore, since cognition is emergent from the loosely-coupled agent-resource complex system, the larger learning gains of the RR LC and the higher sensitivity of the RR AC would not merely relate to 'superficial' resource-specific knowledge, but also deep conceptual understanding and transfer of learning. To the mindbased perspective on the other hand, resources are not central to cognition. Decoupling should not affect learning as long as students still receive adequate feedback. If anything, the Rp LC should facilitate learning, particularly learning for transfer, because it reduces cognitive load associated with managing the resource, freeing up cognitive capacity for focusing on developing a deep understanding of underlying concepts. Furthermore, the mindbased perspective would expect the Rp AC to be more sensitive to differences in learning gains, especially transfer of learning, because it eliminates construct-irrelevant variance related to managing the resource and resource-specific knowledge irrelevant to a deep understanding of underlying concepts.

Procedure

The study was conducted as part of regular in-class exercise activities of the first year BSc introductory programming course. The course took place across fourteen weeks, during fall semester 2018, from late September until late December. It entailed two weekly hours (i.e. 2x45 minutes) of lectures, two weekly hours of on-site exercises in small groups supervised by student teaching assistants (11 groups with between ca. 15-45 students each), weekly homework in Code Expert, and a final sixty minute summative examination in January 2019. The course is mandatory for first semester Bachelor students in Civil Engineering, in Geospatial Engineering, and in Environmental Engineering. The study activities took place in November 2018 during the second hour (45 minutes) of the on-site small group exercises of course weeks nine and ten, with one week between the learning activities and the assessments. We planned the study activities near the end of the course to ensure that all students were deeply familiar with the Code Expert environment, such that differences in assessment performance between the RR and Rp LCs could not reasonably be attributed to increased familiarity of students in the RR LC with surface features of the Code Expert environment. On the first study day, the students engaged in the learning activity consisting of the five-step tutorial on functional programming, either under RR (compiler active) or Rp (compiler deactivated) conditions. On the second study day, the students sat the assessment, again either under RR or Rp conditions. Time available for both the learning activity and the assessment was thirty minutes. While the students could progress through the five tutorial tasks at their own pace, in the assessment they had precisely ten minutes time available for each of the three tasks. In the week between day one and day two, there were no exercises or other activities related to the topics covered on day one. Figure 2 illustrates the experimental procedure: Group1 participated in the RR LC

and Rp AC; Group2 in the RR LC and RR AC; Group3 in the Rp LC and RR AC; and Group4 in the Rp LC and Rp AC.

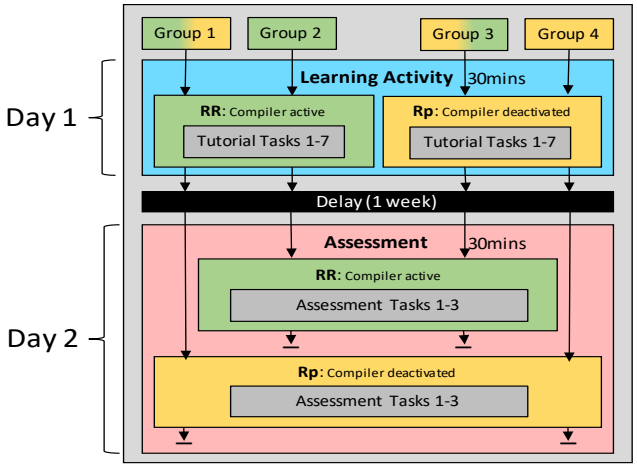

Figure 2: Illustration of the experimental procedure.

Sample

234 out of 272 enrolled students participated in the study. Of these, 21 students participated only on one of the two study days and had to be excluded from the analyses. For another 21 students we could not rule out for certain, that they had not accessed tasks or resources not intended for their experimental conditions, and were also excluded. The resulting sample of n=192 students is distributed across the experimental conditions as follows: n(RR→RR)=49, n(Rp→RR)=48, n(RR→Rp)=53, n(Rp→Rp)=42.

Results

Table 1 reports the mean percentage of points achieved in the complete test consisting of Task1, Task2, and Task3; in the replication task Task1; and in the transfer tasks, Task2 and Task3 taken together. Three things are worth note. First, students in the RR LC outperform students in the Rp one both in the RR and in the Rp assessment and in all tasks. Second, the RR assessment is more difficult (i.e. students performed worse) than the Rp assessment for students of both the Rp *and* the RR LC. Third, the performance difference in the transfer tasks between the two LCs is larger for the Rp assessment, with a 27% difference (63% - 37%) compared to a 14% difference (36% - 22%) in the RR assessment.

Table 1: Mean percentage of points achieved

LC →AC	RR →RR	Rp →RR	RR →Rp	Rp →Rp
Complete Test	46%	31%	70%	47%
Task1	67%	50%	83%	67%
Transfer Tasks	36%	22%	63%	37%

Figure 3 illustrates the assessment results in the complete test, the direct replication task, Task1, and the transfer tasks, Task2 and Task3 together. The vertical histograms illustrate the frequencies (x-axis) of total points achieved (y-axis) for each of the four experimental groups. The background and bar colors represent the LCs and ACs, respectively, green for RR and yellow for Rp. To illustrate appropriate interpretation of the histograms: 67% of students in the RR→Rp experimental condition achieved the maximum of three points in Task1. Furthermore, non-parametric Mann-Whitney *U* inferential statistics, corresponding p-values, effect sizes *r*, and mean ranks (lower values correspond to better performances) are reported for the comparisons between the RR LC and the Rp LC as measured in the RR AC and the Rp AC, respectively. Example: The comparison between the RR and Rp LC as measured by the complete test in the RR AC is highly significant with p=.008, *U*= 1'539, effect size *r*=.27 and better performance of the students from the RR LC (mean rank 41.43 < mean rank 56.42).

All test and subtest comparisons between the RR and Rp LC are statistically significant. The reported effect sizes *r* constitute small to medium effects (Field, 2009). Effect sizes are consistently larger for the Rp test than for the RR test, are consistently larger for the transfer tasks than the direct replication task, and the difference in effect size between RR and Rp assessment is larger for the transfer tasks.

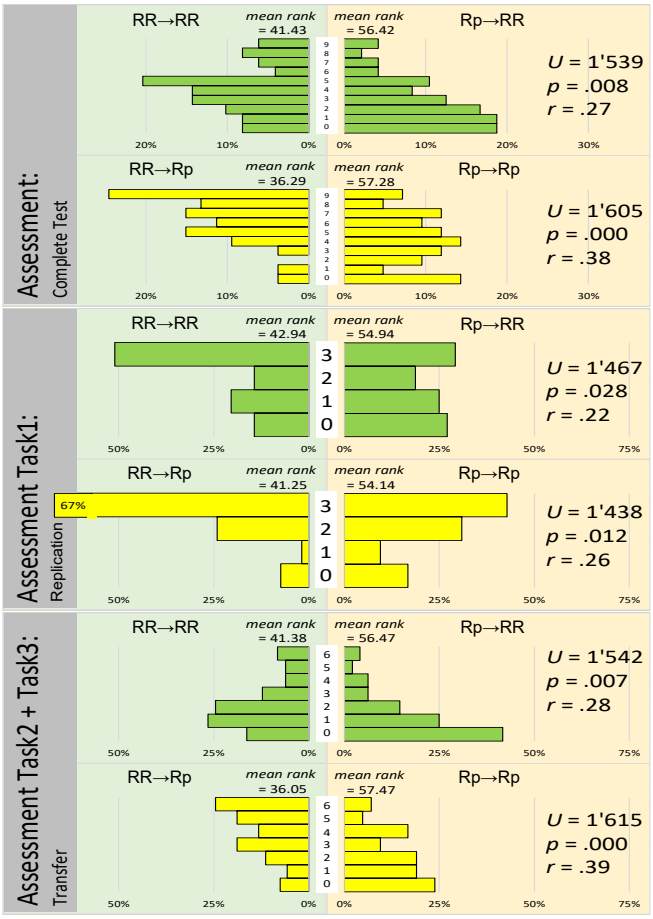

Figure 3: Assessment performances and inferential statistics.

Discussion

Learning

The findings convincingly confirm the proposition that resource mediation facilitates learning. RR learning consistently outperformed Rp learning. Effect sizes were small to medium for the direct replication task and medium for the transfer tasks. Of particular note, the effect is consistent and robust even across RR and Rp ACs, and even after only thirty minutes of tutorial-guided learning. The fact that this effect was stronger in the transfer tasks, and not only in the RR AC but even more so in the Rp AC, is strong evidence that resource-mediation facilitates not just learning of superficial resource-specific details, but in fact deep conceptual understanding and successful learning for transfer. Furthermore, our results support the assumption that it is indeed the presence or absence of practice-relevant resource-affordances that mediated these differences in learning. First, because the only difference between the otherwise identical LCs was whether the compiler was active or not, second, because not only the students in the RR LC, but also the students in the Rp LC received feedback on the correctness of their solution.

Assessment Validity

Both the RR and the Rp assessments successfully differentiate between the two LCs and are thus sensitive to the experimental manipulation of resource-mediated learning. Hence, we cannot reject the validity of neither the RR nor the Rp assessment. However, the Rp assessment was more sensitive to the experimental manipulation than the RR one, and especially in the transfer tasks. We identify two possible reasons for this. First, the higher sensitivity of the Rp assessment could be an indicator of better validity of the Rp assessment in general. Alternatively, the higher sensitivity could be an indicator of superior *differential* validity of the Rp assessment for assessing the transfer of (resource-mediated) learning in specific, but not necessarily for learning outcomes as they relate to the disciplinary practice at large. Two observations support this second interpretation. First, the RR assessment was consistently more difficult than the Rp assessment. It clearly required students to demonstrate competencies that go beyond what would have been sufficient to succeed in the Rp assessment. Second, the RR assessment is more directly representative of the target disciplinary practice of programming (which also includes a functional compiler), i.e. it is more 'ecologically valid'. If we assume that disciplinary competencies in all their complexity usually constitute the intended measurement constructs of examinations, then – somewhat paradoxically – the higher sensitivity observed in this study would imply that the Rp assessment suffers from construct underrepresentation in relation to the ecologically valid intended measurement construct. To further illustrate this argument: If the RR and Rp assessments captured the exact same amounts of variance in *transfer of resource-mediated learning*, but the RR assessment in addition also captured *other* variance relevant to competency in the disciplinary practice, then we would indeed expect precisely the observed pattern of higher sensitivity of the Rp assessment to the experimental manipulation. Taken together, this indicates that RR tasks may render more valid estimates of students' effective competency in a target practice, while Rp tasks may be more valid for the differential assessment of associated (developing) conceptual understanding.

Cognition

Neither the proposed situated nor mindbased perspective facilitated the prediction of the study's results. The mindbased perspective proved rather unsuitable for explaining the substantial and robust learning facilitation in the RR LC, while the situated perspective does not offer a meaningful account for the higher sensitivity of the Rp assessment. Regarding the ontological question of the nature of cognition, it is indeed quite intriguing that the uncoupled 'mindbound' Rp assessment was *more* sensitive to transfer of resource-mediated 'situated' learning, than the RR assessment. This pattern in many ways appears reminiscent of learning as (resource) internalization in a Vygotskian or Piagetian sense. Alternatively, from a complex systems perspective we might conclude that the mindbased perspective does not adequately account for cognitive phenomena emergent from agent-resource interaction, while the situated perspective does not adequately account for near decomposability. Such considerations notwithstanding, our data show that there is something more complex going on than can be explained with either a rigidly mindbound or rigidly situated perspective alone. This approach did not lead to parsimony, but instead to poor predictions.

Implications for Practice

We identify three main implications for practice. First, the RR tasks were more difficult than the Rp ones. When moving from Rp assessments to RR ones, this increase in difficulty needs to be accounted for. We can confirm this experimental finding from our own experience in supporting lecturers when transitioning from conventional Rp paper-based examinations to RR computer-based ones – e.g. with Code Expert. The new RR examinations usually require substantially more time for students to be able to solve them meaningfully. Second, we found robust evidence confirming RR learning activities facilitate deep conceptual learning and successful learning for transfer. If assessment drives learning, then we are well advised to include at least some RR tasks in any examination, providing students an effective incentive to engage in according and productive RR learning activities. Third, mixed examinations consisting of both RR and more conventional Rp tasks may be best, because Rp tasks may be more suitable for the differential assessment of developing conceptual understanding, while RR tasks may be more suitable for 'ecologically valid' and exhaustive assessments of accomplished disciplinary competency.

References

Abrahamsen, A., & Bechtel, W. (2012). History and core themes. In *The Cambridge Handbook of Cognitive Science* (pp. 9–28). Cambridge, UK: Cambridge University Press.

Baird, J.-A., Andrich, D., Hopfenbeck, T. N., & Stobart, G. (2017). Assessment and learning: fields apart? *Assessment in Education: Principles, Policy & Practice*, *24*(3), 317–350.

Bennett, R. E. (2015). The Changing Nature of Educational Assessment. *Review of Research in Education*, *39*(1), 370–407.

Biggs, J. (2014). Constructive alignment in university teaching. *HERDSA Review of Higher Education*, *1*(1), 5–22.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The Concept of Validity. *Psychological Review*, *111*(4), 1061–1071.

Clark, A. (2012). Embodied, embedded, and extended cognition. In *The Cambridge Handbook of Cognitive Science* (pp. 275–291). Cambridge: Cambridge University Press.

Crisp, G., Guàrdia, L., & Hillier, M. (2016). Using e-Assessment to enhance student learning and evidence learning outcomes. *International Journal of Educational Technology in Higher Education*, *13*(1).

Danish, J. A., & Gresalfi, M. (2018). Cognitive and Sociocultural Perspectives on Learning: Tensions and Synergy in the Learning Sciences. In *International Handbook of the Learning Sciences* (pp. 34–43). New York, NY: Routledge.

Field, A. P. (2009). *Discovering statistics using SPSS: and sex, drugs and rock 'n' roll* (3rd ed). Los Angeles: SAGE Publications.

Gibson, J. J. (1977). The theory of affordances. In *Perceiving, acting, and knowing: Toward an ecological psychology* (pp. 67–82). Hillsdale, NJ: Erlbaum.

Gulikers, J. T. M., Bastiaens, T. J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development*, *52*(3), 67–86.

Halbherr, T., Dittmann-Domenichini, N., Piendl, T., & Schlienger, C. (2016). Authentische, kompetenzorientierte Online-Prüfungen an der ETH Zürich. *Zeitschrift für Hochschulentwicklung*, *11*(2), 247–269.

Halbherr, T., Reuter, K., Schneider, D., Schlienger, C., & Piendl, T. (2014). Making Examinations more Valid, Meaningful, and Motivating: The Online Exams Service at ETH Zurich. *EUNIS Journal of Higher Education.*, *1*(1).

Hmelo-Silver, C. E., Kapur, M., & Hamstra, M. (2018). Learning Through Problem Solving. In *International Handbook of the Learning Sciences* (pp. 210–220). New York, NY: Routledge.

Hutchins, E. (1995). How a Cockpit Remembers Its Speeds. *Cognitive Science*, *19*(3), 265–288.

Kane, M. T. (2006). Validation. In *Educational Measurement* (pp. 17–64). Washington, D.C.: American Council on Education.

Keehner, M., Gorin, J. S., Feng, G., & Katz, I. R. (2017). Developing and Validating Cognitive Models in Assessment. In *The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications*. John Wiley & Sons.

Lehner, H., Avanthay, D., & Sichau, D. (2018). *Code Expert*. Retrieved from https://code-expert.net/

Messick, S. (1989). Validity. In *Educational Measurement* (pp. 13–100). Washington, D.C.: American Council on Education.

Mislevy, R. J. (2018). A Sociocognitive Perspective. In *Sociocognitive Foundations of Educational Measurement* (pp. 21–45). New York, NY: Routledge.

Moss, P. A. (1992). Shifting Conceptions of Validity in Educational Measurement: Implications for Performance Assessment. *Review of Educational Research*, *62*(3), 229–258.

Pellegrino, J. (2002). Knowing What Students Know. *Issues in Science and Technology*, *19*(2), 48–52.

Ruch, G. M. (1924). *The improvement of the written examination*. Oxford, England: Scott, Foresman & Co.

Sambell, K., McDowell, L., & Brown, S. (1997). "But is it fair?": An exploratory study of student perceptions of the consequential validity of assessment. *Studies in Educational Evaluation*, *23*(4), 349–371.

Sawyer, R. K. (2014). The New Science of Learning. In *The Cambridge Handbook of the Learning Sciences*. New York: Cambridge University Press.

Schwartz, D. L., & Martin, T. (2004). Inventing to Prepare for Future Learning: The Hidden Efficiency of Encouraging Original Student Production in Statistics Instruction. *Cognition and Instruction*, *22*(2), 129–184.

Shapiro, L. (2011). Standard Cognitive Science. In *New Problems of Philosophy*. *Embodied Cognition* (pp. 7–27). New York: Routledge.