Designing Digital Peer Assessment for Second Language Learning in Low Resource Learning Settings

Maletsabisa Molapo
IBM Research
Johannesburg, South Africa
maletsabisa.molapo@ibm.com

Chane Simone Moodley
IBM Research, Wits University
Johannesburg, South Africa
chane.simone.moodley1@ibm.com

Ismail Yunus Akhalwaya
IBM Research, Wits University
Johannesburg, South Africa
ismaila@za.ibm.com

Toby Kurien
IBM Research
Johannesburg, South Africa
toby.kurien@za.ibm.com

Jay Kloppenberg
African School for Excellence
Johannesburg, South Africa
jay.kloppenberg@ase.org.za

Richard Young
IBM Research
Johannesburg, South Africa
richard.young2@ibm.com

ABSTRACT
In low-resource, over-burdened schools and learning centres, peer assessment systems promise significant practical and pedagogical benefits. Many of the these benefits have been realised in contexts like massive open online courses (MOOCs) and university classrooms which share a specific trait with low-resource schools: high learner-teacher ratios. However, the constraints and considerations for designing and deploying peer assessment systems in low-resource classrooms have not been well-researched and understood, especially for high school. In this paper, we present the design of a peer assessment system for second language learning (English as a Second Language) for high school learners in South Africa. We report findings from multiple studies investigating qualitative and quantitative aspects of peer review, as well as the contextual factors that influence the viability of peer assessment systems in these contexts.

Author Keywords
Peer Assessment; Low-Resource Learning; Mobile Learning; Second Language Learning; Qualitative Methods.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous; See http://acm.org/about/class/1998 for the full list of ACM classifiers. This section is required.

L@S '19, June 24–25, 2019, Chicago, IL, USA

ACM ISBN 978-1-4503-6804-9/19/06…$15.00
https://doi.org/10.1145/3330430.3333626

INTRODUCTION
Proficiency in a wide range of skills is achieved through deliberate practice [21], wherein learning is enabled by repetitive practice, receiving immediate feedback and applying corrections within short cycles. Studies have further shown that if corrective feedback is delayed, the benefit of the feedback is significantly reduced [28,29]. Language learning is one area for which deliberate practice is necessary for learners to achieve proficiency, but in learning settings where learner-teacher ratios are high, it is impractical for instructors to administer frequent writing exercises and deliver timeous feedback. These constraints apply to English learning in some over-burdened schools in South Africa.

In other settings where learner-teacher ratios are high (such as large university classrooms and Massive Open Online Courses - MOOCs), peer assessment or peer review has been shown to provide alternative feedback to enable deliberate practice [2,30]. Successful digital peer assessment systems have been deployed in these settings – including for language learning applications [2,10,32]. Based on these results, we sought to design, explore and test a digital peer assessment system to accelerate English language learning in a context with similar class-size challenges, but a different learning setting: a South African high school situated in a low-income urban community.

Many established digital peer assessment approaches were researched and developed in different contexts than the one researched in this paper. In rural and low-income urban areas, schools are faced with challenges that dictate a more complex setup of a digital peer review system to accommodate a multitude of limitations within the learning ecosystem [36]. In addition to the challenging learning environment, we worked with high school learners who were all early English le arners.

We designed and deployed a prototype peer-assessment system; using user studies, co-design sessions, trial deployments, teacher workshops, and iterative development

and testing. We used our prototypes to explore digital peer assessment for this context, and we report on our findings alongside the evolution of the solution we developed. Beyond the results of our case study, analyses and lessons for the broader learning-at-scale community, we see this paper as a guide that teams and schools looking to set up peer assessment systems in resource-constrained learning environments (classrooms, after-school programs, etc.) can use as a starting point.

BACKGROUND AND RELATED WORK

Setting Up Digital Peer Assessment

Peer assessment often has two goals: 1) to help assess the quality of peers' work by providing quantitative scores, ratings or grades; and/or 2) to provide actionable qualitative feedback [29,30]. When setting up a peer review system, then, design questions include: What type of qualitative or quantitative peer feedback should be elicited from the reviewers, how must it be structured? What should the system measure?

In quantitative assessment, peers can provide either ordinal or cardinal assessment [37]. In cardinal systems, peer reviewers return a numerical score for each allocated submission, then the scores per submission are averaged to compute its final score or grade [23,45]. Reviewers rate the work as a whole or are guided by specified assessment criteria. In ordinal systems, peer reviewers return a ranking of a small subset of allocated peer submissions - ranked from strongest to weakest [52], essentially providing a relative measure of quality for each submission [46]. Reviewers either compare multiple submissions at once: a 'listwise' comparison [53], or two submissions at a time: a 'pairwise' comparison [7]. Beyond the subset size (pairwise or listwise), reviewers compare submissions based on either the whole work (e.g., submission A is stronger than submission B) or on specified criteria (e.g., submission A is stronger than submission B based on criteria X). Using either form of ranking, the output of an ordinal system is a class ranking, based on which various statistical models can be used to estimate the final grade per submission [8,9,26,47]. Some researchers argue that it is easier for junior peer reviewers to provide ordinal feedback than cardinal feedback, since the reviewer is not required to directly score their peers' submissions [24,44]. Cambre et al. also found that when learners were asked to compare submissions instead of providing individual evaluations, they returned reviews of a higher quality [7].

As is the case with quantitative feedback, qualitative feedback can be scaffolded [13] or guided by an itemised rubric. While the use of rubrics has proved beneficial, other researchers found that peer reviewers wrote shorter reviews when rubrics were used [1], so there are open questions around how to effectively use rubrics to encourage high quality qualitative reviews. When choosing among these various options for setting up peer review systems, the objective is to enable peers to submit high quality and educationally beneficial qualitative reviews [25].

Learning At Scale Under Limited Resources

Many learning-at-scale (L@S) solutions are applicable to learning in resource-constrained settings - the main similarity being large classrooms and high learner-teacher ratios. However, major contextual differences exist. In low-resource settings, there may be several infrastructural limitations, learners may originate from complex socio-economic backgrounds, and instructors may be less trained than in developed contexts. Nye [43] studied the barriers to the adoption of digital learning systems in low-resource settings, identifying factors such as ICT hardware availability, electrical and internet reliability, data costs, basic ICT skills, language, and lack of culturally appropriate content.

As we approached this work, we did not see these factors as barriers to adoption but rather as factors that exist in a L@S context not often-studied. Learners should not struggle to adopt a resource-intensive solution developed for a different context than their own [20,35]. Rather, solutions should be designed, from the very beginning, for the resources available in the learning environment, and adapted accordingly to ensure that learners continue to have a pleasant learning experience. The question, then, is how do we design appropriate peer assessment systems within these contextual realities? What technical and non-technical support structures are needed for the productive use of digital peer review in these contexts?

METHODOLOGY

Context

This study was conducted in an independent public school in a low-income urban community in South Africa. The school follows an innovative pedagogical model, optimising on limited human and infrastructural resources. The school was already using peer assessment for English and other subjects before this study, where learners provided feedback on a one-to-one, non-anonymous basis, on paper. The school had about 30 laptop computers at the beginning of the study.

Participants

The learners in the partner school live in the local community. English is not their first language. At various stages of the study, we worked with learners across all grades in the high school (total N=201; our sub-studies had different participation and we report accordingly per study). Of all the learners in the school, 89% owned a smartphone of their own (over 70% being budget smartphones), 63% had access to a smartphone at home or via a close relative, and 2% (n=4) did not have access to a smartphone at all. 90% of the learners' smartphones were Android devices, 1% ran iOS and 9% ran the Tizen OS. Regarding internet access and data costs, 40% of the learners in the study said they purchased monthly WhatsApp bundles (giving them access to WhatsApp only) and used their parents' or siblings' devices if they needed the internet for anything else; 30% bought monthly data bundles

(typically 500MB-1GB a month). The rest said they bought weekly or ad-hoc bundles, including 'night-owl' bundles (significantly cheaper data bundles available only between 00:00 and 06:00). English instructors and teaching assistants (N=5) participated in the study throughout. Author 5 is the co-founder of the partner school and contributed pedagogical expertise and institutional understanding to the design process.

Methods

We spent a total of 15 months on the work reported in this paper. Before this period, we established a relationship with the school, visited and participated in multiple classes, and reviewed related literature on peer review. These led us to build the first iteration of our prototype (functionality described in the next section). Using iterative co-design, we developed three iterations on our peer assessment system, and we present the details of its evolution. During the iterative design process, we conducted multiple studies to explore feedback features, optimise the peer review experience, and adapt to the local context. We conducted two usability studies, two deployment studies, two exploratory studies investigating qualitative and quantitative feedback features, two teachers' workshops, and multiple co-design sessions and semi-structured interviews with both teachers and learners. Qualitative data was analysed using thematic analysis by two researchers.

INITIAL EXPLORATION OF DIGITAL PEER ASSESSMENT

Exploratory Prototype – Version 1

We developed the first prototype of the digital peer assessment system as a starting point to enable explorations with the teachers and learners [38]. The design of the first prototype was based on the literature and initial discussions with the school. The system was a progressive web application that could be viewed on a computer or mobile browser. Teachers could set writing prompts for learners to submit essays and reviews directly on the system. The review process was anonymous. For quantitative feedback, we implemented *pairwise ordinal feedback*. During review, each peer reviewer would be presented with peer submissions in pairs, and requested to compare the submissions and indicate by how much (significantly or slightly) one was seen as stronger than the other – Figure 1 (a). The qualitative feedback was gathered per submission – Figure 1 (b). This was *scaffolded essay-level feedback.* Each reviewer would enter comments in response to a set of three guiding questions: *1) What did you like about this work?, 2) How can the author improve this work? 3) How well did the author do on X?* (X being an assessment criterion specified by the teacher, e.g., "How well did the author *show and not tell?*"

We developed three main algorithms to enable the functionality of the prototype (reported in detail in a separate publication). First, a graph-based pairing algorithm that optimally handed out pairs such that 1) all submissions are marked approximately the same number of times, 2) there are no disconnected clusters of compared pairs, 3) a learner never marks their own submission, 4) a learner never marks the same pair more than once, and 5) a rank-able graph is quickly reachable. In addition, we implemented an algorithm that used the pairwise comparisons to rank the submissions from best to worst. Based on the ranking, the teacher would be prompted to grade the highest and lowest submissions on the ranked list, and these, along with the ranking, would be used to infer grades for the whole class. We also implemented a Bayesian inference system, where the pair-wise comparisons were mathematically modeled taking into account "true" grades and reviewer reliability and variability. Given the learner-provided comparison data, the Bayesian process allowed us to infer "true" grades and individual reviewer reliability. This helped us deal with noisy and sometimes contradictory comparisons. Reviewer reliability was also a useful pedagogical signal that the teacher could use both to incentivise the reviewers and provide feedback on the quality of their reviews. Other advantages of our Bayesian approach include the ability to provide uncertainty measures for the computed rank and grades, and the ability to intelligently hand out pairs to reduce the uncertainty.

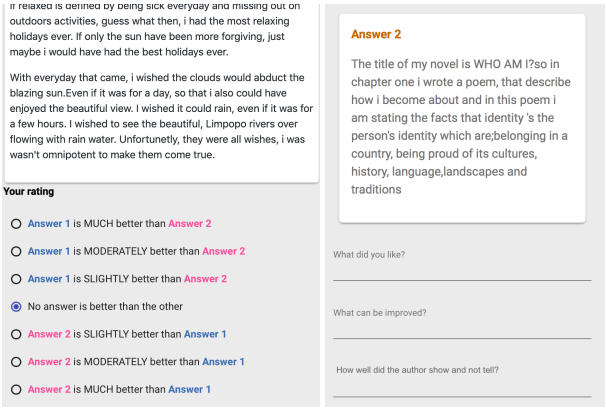

Figure 1 - Reviewing on the v1 prototype: a) Quantitative ordinal feedback, b) Qualitative feedback per submission

Initial Exploration: Study

The first study after developing the prototype was a *usability study*. We conducted an early testing session 18 learners from across three grades (Grade 8-10) who had not been exposed to the prototype before. The learners were assigned a writing exercise, and individually, each participant logged into the system and followed a set of given tasks to submit an essay, then review and rate peer submissions. We observed the interactions, then held a focused group discussion to capture further feedback on the prototype. Following the usability study, the prototype was updated to fix some robustness issues identified in the study, then deployed for four months in the school for further testing in-situ. This was our *first deployment study*. During the deployment, the prototype was used by 111 learners across three grades for English writing classes, generating over 300 submissions. In this period, the research team visited the school four times for observations and semi-structured

interviews with teachers and learners, and to identify evolving needs and preferences.

Initial Exploration: Results

Direct Annotations, Edits and Corrections: For peer review of English writing before we introduced the digital system, the learners used pen on paper to annotate and suggest edits on their peers' essays. So from as early as the first usability study, they requested for a similar functionality on the digital peer review system. One learner's feedback read: *"I would like to show my peers exactly where their mistakes are and how to correct them, instead of just typing a general comment at the end of their essay".* From these direct edits, the authoring learner would clearly see the location of their writing errors, and receive direct recommendations on how to address them.

Resource Limitations: The school did not have enough computers for all the learners to write and review on the system as often as was necessary for repetitive writing and correction. The school had 28-30 functional laptop computers that were shared across multiple grades and used for multiple subjects and activities. So, the learners could only access the writing system about once a week, for 45 minutes, which was not enough to write, review and revise. Due to these limitations, and exacerbated by the learners' slow typing speed, feedback came slower than excepted.

Confusing Whole-Work Comparisons: On the system, the learners were meant to compare two submissions at a time as seen on Figure 1 (a), but it was not trivial for most of them. In many of the feedback sessions, the learners asked: *"Based on what do I say Answer 1 is better than Answer 2? One paper has very good grammar, but the other has better sentence structure. How do I say which is* ***better?***"

Being Deliberate About Deliberate Practice: We learned from reflections with the teaching staff and engagements with the learners that, for a system developed specifically for language learning, it would be important to separate deliberate practice (enabled by qualitative feedback) from quantitative feedback (used only to generate a grade per submission). Instead, the preference was to have a system that would allow the learners to write, receive qualitative feedback from their peers, revise their work and resubmit multiple times before being graded (via pairwise comparisons). The teachers also emphasised *measuring what matters*, stating that to maximise the learning outcomes of the peer review process, the system should enable measuring more than the overall grades per exercise: it must capture feedback in ways that will enable the measurement of more important aspects of writing. For example, they asked: *"Can we collect feedback that would enable us to say, after six months, that learner A is improving in their use of verbs?*

EXPLORING FEEDBACK FRAMING

As necessitated by the findings from our initial explorations, we ran further exploratory studies to co-investigate, with the teachers and the learners, the right features for collecting and framing quantitative and qualitative peer feedback.

Framing Quantitative Feedback – The Study

For quantitative feedback, we investigated input parameters for ordinal grading. This was in response to the learners' frustrations with unguided pairwise comparisons. We ran the exploratory study with 16 participants (13 learners and 3 teachers), investigating:

- Should peer reviewers compare multiple submissions at once (listwise), or compare two submissions at a time (pairwise)?
- Should each reviewer compare submissions based on the submission as a whole (whole-work based comparisons) or on specified assessment criteria (rubric-guided comparisons)?

We selected eight papers from previously submitted essays, representing a combination of weak and strong submissions by teachers' ratings. Each learner reviewed four papers using four different approaches (described below). First, each participant read the four papers allocated to them, and was then given four different ranking sheets to rank the papers – each paper in one of the four ranking approaches. The learners were divided into four groups, and the order of exposure to the ranking approaches was balanced such that each group saw a different approach first, and in the end, every participant was exposed to each of the four approaches.

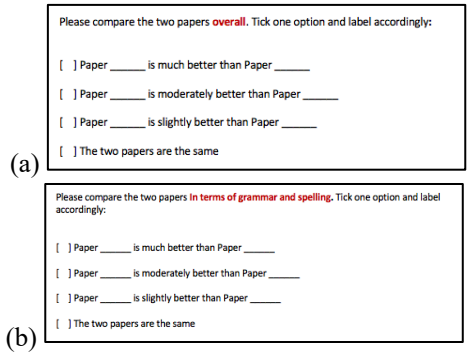

Figure 2 – Mockup of the pairwise comparison: a) Based on the whole work, b) Based on specified assessment criteria

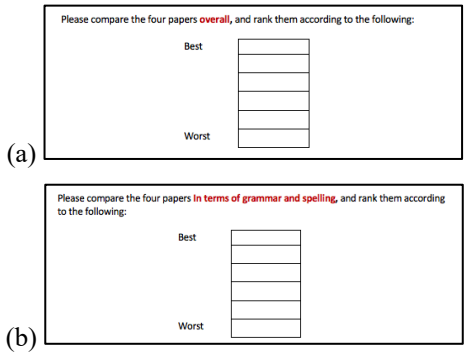

Figure 3 - Mockup of the list comparison: a) Based on the whole work, b) Based on specified assessment criteria

The ranking approaches were: 1) A whole-work-based pairwise comparison, 2) a rubric-based pairwise comparison, 3) a whole-work-based listwise comparison, and 4) a rubric-based listwise comparison (Illustrated in figures 2 and 3). For the listwise comparisons, the ranking form had six slots for placing the four papers in the perceived order of quality, with two extra spaces to enable the reviewer to show the relative distance between the papers. Examples of filled forms are shown on Figure 4. The ranking sheets were low fidelity prototypes that allowed us to test the four comparison approaches without developing them as features on the system.

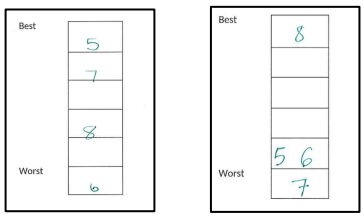

Figure 4 - Examples of filled listwise forms by two different reviewers - showing relative perceptions of quality

The teacher participants were asked to manually grade the papers used in the study, and their marks were averaged to generate a ground truth ranking. All the participants (including the teachers) then rated the four ordinal grading approaches by preference, which was followed by a focused group discussion.

Framing Quantitative Feedback – Results

Rubric-Based Listwise Comparison Was Preferred The Most: 75% of the participants (12 of 16) preferred the rubric-guided listwise approach over all others. The other 25% preferred the rubric-guided pairwise comparison. The participants preferred listwise comparisons (over pairwise) because it gives the reviewer a broader view of the subset of papers allocated to them, enabling one to make a clear decision on the relative comparison of all – as opposed to seeing two at a time and never having an opportunity to compare disconnected pairs. For clarity, they preferred rubric guided comparisons over whole-work based comparisons because they felt that ranking papers based on one criterion or a set of criteria would be more specific and guided than a whole-work comparison. In particular, the teachers felt that generating rankings based on assessment criteria would enable the system to track the performance of each learner against different assessment criteria over time – an additional data point for meaningful measurement.

Both Pairwise and Listwise Comparisons Generated Credible Rankings: The outcome of an ordinal assessment is a class ranking generated from the individual reviewers' rankings or comparison inputs. So, in this study, we compared the ground truth ranking (teachers' ranking) against those that resulted from the listwise and the pairwise comparisons (looking at whole work cases since the teachers provided a whole work grading and ranking). We calculated a Kendall's tau distance to determine the disagreements between the ground truth ranking and the listwise vs pairwise rankings. The Kendall's tau distance is a measure of the correspondence between two rankings; where values close to 1 indicate strong agreement and values close to -1 indicate strong disagreement. We found a Kendall's tau distance of $K=0.49$ for ground truth ranking vs pairwise ranking, and $K=0.57$ for ground truth ranking vs listwise ranking. So in terms of the accuracy of listwise vs pairwise rankings, listwise ranking yielded a slightly higher agreement with the ground truth but both were in close agreement (Figures 5, 6).

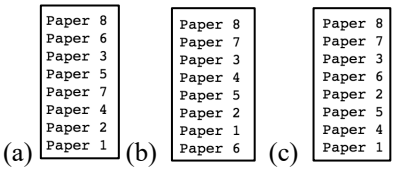

Figure 5 - a) Ground Truth Ranking, (b) Pairwise Comparison Ranking, (c) Listwise Comparison Ranking

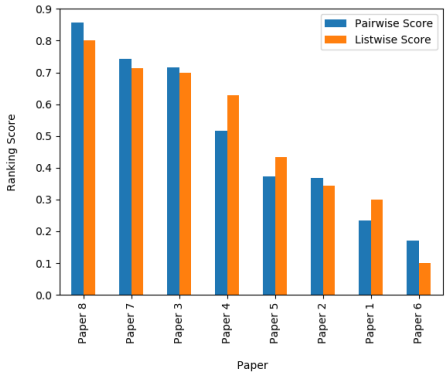

Figure 6 - Paper ranking scores, comparing pairwise and listwise scores (whole work ranking)

Framing Qualitative Feedback – The Study

For qualitative feedback, we investigated open ended essay-level comments and direct inline annotations and edits, asking:

- Does direct annotation and editing enable peer reviewers to generate useful, actionable feedback that is not easily captured using essay-level comments?
- With the direct annotation feature on the system, are open-ended, essay level comments still necessary?

For this study, we worked with 37 Grade 10 learners who already had experience with the prototype peer assessment system. Each of the learners wrote an essay responding to a writing prompt assigned by their English teacher. Each essay was then reviewed by two anonymous peers (focusing only on qualitative feedback and revision), after which the learners re-wrote the essay using the feedback they received. The essays were written and reviewed on the same day, and the second drafts written the next day. The class was divided into three groups where each group focused on one type of feedback. Group 1: word-level or sentence-level

annotation/inline editing only; Group 2: only essay-level scaffolded comments in response to the questions: *"what did you like about the paper?", "what can the author improve?",* and *"comment on the author's creativity in this submission"*; and Group 3: both word-level or sentence-level annotation/inline editing and scaffolded essay-level comments. Our objective was to study the quality, usefulness and actionability of the feedback generated by the three approaches. Since each paper was reviewed twice, this means we analysed 74 reviews: 25 reviewed by Group 1, 24 reviewed by Group 2, and 25 reviewed by Group 3. No learner reviewed their own submission or reviewed two submissions from the same author. Each submission was reviewed using two of three approaches.

Framing Qualitative Feedback – Results

Annotation and Essay-Level Comments Are Complimentary: We found that in papers where reviewers only provided word-level and sentence-level (inline) annotations as feedback, the review lacked a holistic view; and that essay-level comments, when used alone without direct annotation and word-level or sentence-level comments, generated non-specific recommendations. Direct annotation enabled reviewers to directly highlight and correct spelling, grammatical and other types of errors - and most authors who were given corrections effected the suggested changes in their second drafts. On the other hand, essay-level comments enabled reviewers to present valuable feedback on various aspects of the overall structure of the essay and on repeating mistakes – which were not present in annotation-only feedback. For example: *"your essay has to be more cohesive,*" or "*in general, you tend to write long sentences."* From the collaborative reflection with the teachers, we concluded that annotations/edits and essay-level comments are needed in combination as they enable complimentary feedback. Direct editing enables more specific suggestions but essay-level comments enable reflection on the work as a whole. However, we found that there is still further work to be done to encourage the reviewers to reflect deeper on the overall quality of their peers' work, and to write more meaningful reviews.

Overlooked Errors: As can be expected from reviewers who are, themselves, early language learners, we found that in submissions reviewed using both inline editing and essay-level commenting, a significant number of writing mistakes were overlooked by the reviewers – 84% of the papers had at least three mistakes that were not highlighted. This was more apparent in inline annotation and editing, because the reviewers had the opportunity to point directly to the errors.

Incorrect Corrections: We also found that in about 22% of the essays on which reviewers gave annotated feedback, the reviewers made wrong suggestions and corrections. The most common wrong corrections were attempts at spelling corrections, where misspelt words (and in five cases, correctly spelt words) were corrected to a wrong spelling.

Reviewers Do Not Want to Correct One Error Repeatedly: The learners requested a future feature where, once an error has been flagged by the reviewer, the system identifies and highlights all occurrences of the same error, e.g., a consistently misspelled word. Without this, they felt obligated to correct all the occurrences of the same error.

Iteration By Itself Enables Improvement: Confirming findings from the literature [22], we found that in the second drafts, the learners made significant improvements, even making multiple changes that none of their reviewers had highlighted. This confirmed that iterating and practising does help one improve their draft, and that exposure to the work of others is also a learning opportunity.

Some Aspects of Writing Are Assessed Better Quantitatively: The learners reported that for rubric guided feedback, it was easier to quantify than describe their assessment. Case in point, in response to the question *"comment on the author's creativity,"* up to 12 learners, without being prompted, returned ratings of the submission's creativity (e.g., 3/5, 3 stars) instead of textual comments on the creativity of the submission. We found that essay-level comments are best given as whole-work assessments, that direct annotations are best for pointing out specific word-level and sentence level feedback; but rubric-based assessments are better off rated cardinally or assessed using ordinal comparisons. This confirmed our findings from the quantitative feedback study, where the results showed that for ordinal feedback, the learners preferred to compare or rate based on specified rubrics, and that for the whole-work, they preferred to give overall comments or text-specific annotations and edits.

Characterising Inline Annotations And Edits: For inline annotation and edits, about 70% of the errors highlighted were spelling and grammatical errors, and among them 20% did not include a suggestion or correction. In some cases (24 of the 50 reviews where inline editing was done), students rewrote a full sentence for the author to show them how the sentence would have been phrased. Other types of inline comments were about good writing practices where the reviewers highlighted a sentence and added a comment such as, *"Avoid long sentences without punctuation,"* or *"It is not recommended to write in the first person."*

ADDRESSING CONTEXTUAL CHALLENGES, REDESIGNING FEEDBACK FEATURES

Supporting Mobile and Offline Access – Version 2

First, to address the resource limitations identified in the exploratory stage, the research team and the school looked at various alternatives, some of which (e.g., providing more computers) were not feasible. It was, however, noted that most students owned smartphones. To take advantage of the learners' devices for writing and peer assessment, we further optimised the peer assessment system for mobile access, specifically for slow, low-end smartphones. For example, the first prototype used Google login, which was fast on a high-end device on a high-speed network, but could take over 15

minutes to confirm the Google permissions on a slow device and a slow network. The login process was updated to include local application login (which was faster) with the option of using Google login for those with faster devices.

However, writing long essays on a small, on-screen keyboard would be difficult. So we explored the use of physical keyboards to be used alongside the phones for writing. At first, we tried wired keyboards which connect to mobile devices using On-The-Go (OTG) cables (each wired keyboard cost about 4 USD), but almost all the learners' low-end smartphones did not support OTG, so we explored Bluetooth keyboards, and these worked for all phone types except Tizen-OS phones. Research funds enabled us to donate 90 physical Bluetooth keyboards to the school. Each keyboard cost about 20 USD. Physical keyboards allowed the learners to view their entire phone screen when writing (Figure 7). The keyboards would be shared among all the learners in the junior school (~120 learners).

Figure 7 – A learner using a physical Bluetooth keyboard for writing

In version 2, we also implemented an offline access feature to enable access to the system at home. Writing prompts and any drafts composed at school were now available offline, and any writing or reviews completed offline would be automatically uploaded when the device went back online.

Redesigning Feedback Features – Version 3

The studies on framing qualitative and quantitative feedback allowed us to determine key areas that would improve how feedback was captured on the system.

On quantitative feedback: In version 3, ordinal grading was updated such that comparisons were guided by specified assessment criteria instead of learners comparing essays on the whole work. This meant that the teacher could set the rubric (assessment criteria) based on which peer reviewers were to rank or compare essays. In future versions, the system will be upgraded from pairwise to listwise comparisons, as guided by the results of our exploratory study on quantitative feedback.

On qualitative feedback, we fully implemented direct editing and annotation. When designing the annotation feature, the teachers suggested that inline edits and annotations generally fall under one of two broad categories: Praise and Improve. "Feedback as praise" often compliments or reinforces good writing practices. "Feedback to improve" includes identifying the error type (e.g., *spelling error*), and then suggesting a correction; or in some cases just highlighting a confusing or unclear statement without necessarily offering a suggestion. This was implemented by allowing the reviewer to highlight a word or sentence, choose the feedback category (Praise or Improve), and leave a comment or correction: Figure 8 (a-c).

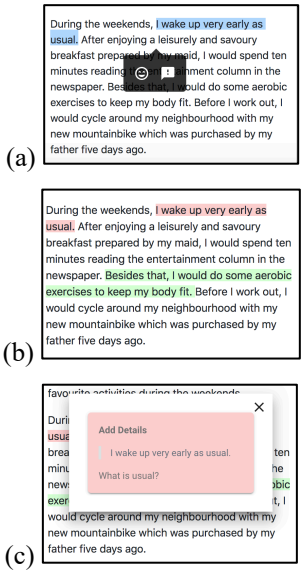

Figure 8 - Inline editing and annotation on mobile. (a) Options given to the reviewer after highlighting text: icons represent praise and improve, (b) How the essay looks after editing. Clicking on the annotation allows the author to see the annotated feedback, as in (c).

On Deliberate Practice: In version 3, we separated deliberate practice from ordinal grading. In this version, once a writing exercise had been published, learners had the option to take the exercise and submit drafts as often as they wished. They could revise and re-submit one draft after another, and their latest draft in the system would be the one that got assigned to the next available peer reviewer. During the *"write, review, revise"* mode, peer reviewers could provide their qualitative reviews in the form of direct edits and/or overall comments. Authors could receive peer feedback from multiple peers and use it to revise their latest draft. If desired, teachers also had an option to provide the qualitative feedback. When all learners had had the chance to write, review and revise, the teacher could then change the system mode to "rating by comparison": where learners used the rubric-guided ordinal approach to compare submissions and generate grades for the writing exercise.

Deploying Version 2 and 3 – The Study

After the new features had been developed and tested, we deployed version 2 of the system (with mobile optimisation and offline access) and the physical keyboards. This was our

second deployment study, which ran for two months, leading up to school holidays, at the end of which we deployed version 3 (with the new feedback features) – our *third deployment study*, which ran for three months. When launching both deployments, we provided initial training for the learners and the teachers on the new features and/or hardware, observed a live exercise using the new features, followed by an in-depth discussion with the learners and teachers. During the deployments, we visited the school twice a week to provide technical support, observe, and co-design solutions to ongoing challenges with the learners and the teachers. At the end of the third deployment, we conducted a school-wide survey with 161 learners to understand their perceptions and experiences with the peer assessment system. During the deployment of Version 3 (3 months), 193 learners took at least one exercise, 1280 drafts were submitted, and 2020 reviews submitted. Of these learners, 70% used their phones to access the system at least once. .

Deploying Version 2 and 3 – Results

Tedious Mobile Set Up: Since the keyboards were a shared resource, the learners had to pair them to their mobile devices before every writing class, then unpair and return them afterwards. Simultaneous attempts to pair with over 30 Bluetooth keyboards of the same Bluetooth name (the keyboards could not be renamed) became a long and painful process for each class. In addition, the learners had to connect to a specific Wi-Fi network enabled for mobile use, which took time. Once connected, accessing and logging onto the peer assessment system, too, took time - webpages loaded slowly on devices with low processing speeds. On average, it took over 30 minutes for a full class to be fully connected and set up, leaving less than 15 minutes for the writing exercise. This is the reality where there are no readily available computer labs. Over time, we collaboratively brainstormed solutions to reduce the setup time; for example, permanently assigning Bluetooth keyboards to individual learners so that they could be permanently paired to their respective phones. While this helped the ~90 learners who received the keyboards (Grades 8,9), the rest of the school could not use the keyboards nor have regular access to the system.

Learners' Perspectives on Writing on Mobile Devices: During writing and review sessions, the teachers provided school laptops to the learners who did not have smartphones of their own or those whose phones did not have the right specifications to run the system (e.g., too slow or insufficient storage capacity to install the minimum version of the Google Chrome browser required by the system). While the learners were generally enthusiastic about using Bluetooth keyboards (especially for writing and typing practice at home), we later learned that those who used their own phones in the classroom felt disadvantaged because the experience of writing on a smartphone was worse than writing on a laptop. The learners were also discouraged by the frequent connectivity issues and the time it took to get the mobile set-up running. Over time, several learners began to deliberately leave their phones or assigned keyboards at home so that they, too, could be assigned laptops in the classroom – which was problematic because the laptops were not enough for everyone.

Learners' Perspectives on Feedback and Learning: Almost 85% of the 161 learners who were surveyed agreed that the peer assessment system improved their English learning experience, and that they benefitted from being exposed to their peers' writing and thinking. About 50% explicitly stated that the system enabled them to improve their typing skills, writing styles, grammar, and spelling. Approximately 46% noted that the direct inline-editing feature assisted them the most – citing it as the most enjoyable feature and stating that it improved the quality of the feedback by making mistakes clearly visible. On the other hand, confirming what we found in our study on qualitative feedback, 24% of the learners mentioned that they often received incorrect reviews and wrong corrections. They suggested that the system could 1) automatically detect and correct incorrect reviews, 2) enable authors to indicate incorrect corrections, and 3) enable the peers to debate suggested corrections on the platform.

Teachers' Experiences, Teachers as Gate-Keepers: We observed that technical glitches like the intermittent network, learners forgetting their passwords, errors due to phone limitations, etc., frustrated most of the teachers and affected their zeal to use the system. However, our regular presence at the school enabled us to attend to and provide support on these issues, and to maintain a strong relationship with the teachers. As we continually solved emerging contextual and technical challenges together, we saw an increase in the use of the system for writing and reviewing on a weekly basis. This confirms previous research that showed that regular engagement with the research/technical team is an enabler for the continued use of a technological intervention [39]; and that a sustainable intervention will require identifying and training technical experts and/or champions within the local community [3,34].

DISCUSSION

Annotation as Qualitative Feedback and Extra Review Data

For a learner to benefit from the feedback they receive, feedback must be specific and immediately actionable [22]. In many peer review systems, comments (often scaffolded) are seen as the main type of qualitative feedback. Through our studies, we found that direct editing and annotation on the submission opens new possibilities for feedback generation and cross-peer learning. It also creates new 'peer correction' data for Natural Language Processing (NLP) - based models to support the peer assessment process. For example, a lot can be learned from the correct and incorrect annotations and corrections made by novice peer reviewers, considering locations on the submissions where annotations were made. Even for the author's benefit, we found that for

learners who are novices at the subject matter being studied (such as, in our case, early English writers), it may not be enough for an author to receive generic comments without direct in-line suggestions. However, as demonstrated by our results, annotated word-level and sentence-level feedback does not remove the need for essay-level feedback, since they are complimentary.

Early Language Learners as Peer Reviewers – Correcting Reviews and Encouraging High Quality Reviews

In our studies, we found that, as expected, early language learners are likely to give incorrect peer feedback, with cases of many inaccurate annotated corrections. However, peer feedback can still be valuable to both the reviewer and the author [10,12], and various NLP techniques can be used to help early language learners to write more accurate reviews. For example, Nguyen et al. [42] developed algorithms to detect and alert reviewers of low-quality reviews, and Kulkarni et al.'s system [29] suggested improvements to the reviews during composition. Similar approaches can be used specifically for early language learners serving as reviewers.

Related to incorrect annotated feedback, our results showed that learners need to be guided and encouraged to write high quality reviews – especially essay-level comments. There is active research from various contexts on how to address this issue. Kulkarni et al. [30] compare peer assessment to crowdsourcing where crowd workers perform better when they are intrinsically motivated by the task's importance. Other researchers have found that awarding good reviewers (where learners are graded on their submissions and their reviews) provided this motivation [11,49]. We seek to investigate this further in our context as part of future work.

Peer and Automated Language Assessment

Our results show that there is need (and potential) to improve the peer review process for early language learning by combining both peer assessment and automated language assessment. This combined approach would leverage the human element of the peer which of is absent in completely automated writing assessment systems [51], while computational intelligence would help augment the peer review process. In our future work, we will experiment with and extend current NLP and computational linguistics approaches for supporting language learning [31]; including error detection [19,48], error correction [14–16], review summarization [33,54], automated writing assessment [6,50,55], learner proficiency classification [17,18], and developing adaptive features to support learning at the point of writing, and when composing reviews. Additionally, our future work will model the influence of previously learned African languages on errors made in early English writing.

Setting Up Peer Review Systems In Low-Resource Learning Settings

Due to resource limitations (limited computers, internet access, electricity, etc.) in many low-resource learning settings, peer assessment systems cannot be set up and used similarly to developed contexts. In many places, mobile phones will for a long time, be the only opportunity for learners to access peer review on a digital platform. First, this means that the system may have to be reimagined as a collaborative homework or afterschool tool, not just an in-classroom tool. It also means that schools have to rethink mobile phone policies, and find ways to manage mobile phones in schools while learners use them to access educational tools. However, new challenges come with the use of mobile phones for educational purposes. First, data costs, both to the school and the learners. Even if there is reliable Internet access at school (which is rare), learners do not have Internet access at home. A system that works offline is therefore ideal: one where learners can write and review at home, then sync to the system when they return to school (whether the system at school is deployed to the cloud or to a local server that is not connected to the Internet).

In addition, while learners may have smartphones, these are often low end smartphones with limited storage, memory and low processing speeds. This limits the types of applications that can be developed for these contexts [41], and necessitates extensive testing to ensure that the peer assessment system is optimised for all learner devices. Moreover, entering an essay, or reviewing one, on a mobile phone is not easy due to the small phone screen size and the onscreen keyboard. We have imagined two solutions to this problem: 1) to use a physical keyboard to connect to the phone and use for typing, or 2) to use Optical Character Recognition (OCR) to scan hand-written work onto the system using mobile devices, such that the bulk of the writing is done by hand and only computations and certain reviews are done on the mobile device. In our work, we tried option 1), but for our deployments at scale where procuring physical keyboards will not be possible, we are exploring paper-phone integration using OCR.

With these realities, the question might be: is it worth it to pursue digital systems for peer review in low-resource learning settings? Our view, based on our work, is: if we can manage to make peer review work within the reality of the learning environment, a digital service (over a purely paper-based system) is still valuable because of the learning possibilities it enables and accelerates – such as using powerful NLP techniques to solve unique learning challenges. These functionalities, which support learning in environments where access to quality teachers is scarce, make it worth it to pursue appropriately designed peer review systems. Lastly, while it is true in most educational settings, this is more so in low-resource settings: the teacher is not always right. In fact, studies in low-resource schools in South Africa have found that teachers are often not highly proficient in the core subjects (e.g., in English) [27]. So we have to ask: how do we design peer assessment systems that allow learners to go beyond and discover new ways of thinking and learn new concepts beyond the teacher and the classroom? Using digital systems allows us to use data to improve learning beyond the limitations of the learning environment, and in ways that would not be possible

otherwise. Remaining aware, as indicated by previous research [5,40] and through the work we reported in this paper, that this means looking beyond the typical: to design for device sharing, for offline access, for mobile-first access, for paper-computer integration, and for new methods of engagement.

Digital Peer Assessment At Scale

It is well-understood that an educational technology solution, in isolation, does not determine the true impact of the intervention [4]. Various social, human and contextual factors influence the real impact of technologies in learning environments; and in low resource settings, these factors require even closer attention [20]. Through our work, we found that a real and impactful deployment of a digital peer assessment system in low-resource learning settings requires deliberate effort to ensure that:

1) The school or learning centre has a well-defined structure for writing and peer review, i.e., when will the learners write, review, etc.,

2) There is teacher buy-in, which means the system has to solve a problem that matters for the teachers, and the benefits of its use must outweigh the efforts to learn it and set it up,

3) The school has enough computers for all learners to write and review. If there are no computers, then there must be enough mobile devices (phones, tablets) with physical keyboards if the system will be used for elaborate writing – or a working combination of mobile devices, computers, and paper,

4) The system is completely available offline, is locally hosted, or the school has a functional internet connection: which means both broadband coverage and data affordability,

5) There is regular technical support. In any context, teachers should be not in charge of overwhelming technical responsibilities that go beyond their regular teaching responsibilities, and

6) It is clearly understood how feedback will be framed and elicited on the peer review system according to the specific learning setting.

Selecting Feedback Input for Peer Review

Based on the various studies we ran to explore how to structure feedback input, we present this generic guide, that can guide any team exploring options for designing the right feedback input features. The goal is to optimise feedback input such that the reviewer gives enough information to meet the desired peer review outcomes, enabling thoughtful and useful qualitative feedback and/or accurate quantitative feedback, yet ensuring that there is no duplication of effort and unnecessary overhead for the reviewer.

First, the outcomes of the peer review process must be determined, as these outcomes will inform what information to seek from the peer reviewer, and in what format. The next step is to decide which input or feedback from the reviewer, when used alone or when computed upon (numerically, statistically or analysed as text) will enable the desired outcome. Specifically, for each outcome, choices involve figuring out: 1) Will these outcomes require qualitative feedback from the peer reviewers, quantitative measurements, or both? 2) For both qualitative and quantitative input, will the reviewer offer this feedback based on the whole work or will a rubric be used to guide them? There are different types of qualitative and quantitative feedback as seen in the literature, that must be chosen based on the desired outcomes.

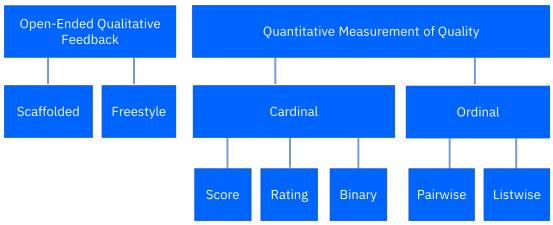

Figure 9 - Variations of Peer Feedback

Qualitative Feedback	Quantitative Measurement	
Provide your feedback on this paper What did you like about this paper?	Rate this paper out of 100 Compare this paper vs this paper	Whole Work
Comment on this paper's sentence construction How good is this paper's sentence construction?	Rate this paper's sentence construction out of 100 Which paper (paper 1 vs paper 2) is better in terms of sentence construction	Specified Criteria

Figure 10 - Examples of qualitative and quantitative feedback.

In our work, we aimed to generate actionable qualitative feedback that addressed specific and holistic aspects of writing, to enable deliberate practice and meaningful short-term and long-term measurements of proficiency. To achieve this, qualitative feedback comprised of whole-work essay-level comments, annotated word-level and sentence-level feedback, and rubric-guided ordinal measurement.

CONCLUSION

Previous research has shown the potential of deliberate practice to improve writing skills when one is learning a new language. We leveraged research that shows that peer review is a viable option to give learners an opportunity for deliberate practice, and co-designed a mobile-accessible peer assessment system for second language learning, working with a South African high school. We presented our process of designing this tool, highlighting our key results relating to the technical design of the system and the contextual realities around which we had to adapt it. We discussed our results and the lessons that are applicable to the wider L@S community, then specifically reflected on and discussed the viability of digital peer assessment systems in low resource learning settings. Our future work includes a grand assessment of the efficacy of a sufficiently-mature version of our solution at broader scale.

REFERENCES

1. Luca de Alfaro and Michael Shavlovsky. 2013. CrowdGrader: Crowdsourcing the Evaluation of Homework Assignments. https://doi.org/10.1145/2538862.2538900
2. Stephen Balfour. 2013. Assessing Writing in MOOCs: Automated Essay Scoring and Calibrated Peer Review (tm). *Research & Practice in Assessment* 8.
3. Edwin H. Blake and William D. Tucker. 2006. User interfaces for communication bridges across the digital divide. *AI & SOCIETY* 20, 2: 232–242. https://doi.org/10.1007/s00146-005-0018-1
4. Bridges.org. 2005. Real Access / Real Impact criteria | bridges.org.
5. Emma Brunskill, Sunil Garg, Clint Tseng, U Washington, and Leah Findlater. Evaluating an Adaptive Multi-User Educational Tool for Low-Resource Environments. *Test*.
6. Aoife Cahill. *Deep Learning for Automated Scoring* *. Retrieved October 17, 2018 from www.kaggle.com/c/asap-aes
7. Julia Cambre, Scott Klemmer, and Chinmay Kulkarni. 2018. Juxtapeer:Comparative Peer Review Yields Higher Quality Feedback and Promotes Deeper Reflection. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 1–13. https://doi.org/10.1145/3173574.3173868
8. Nicola Capuano, Vincenzo Loia, and Francesco Orciuoli. 2017. A Fuzzy Group Decision Making Model for Ordinal Peer Assessment. *IEEE Transactions on Learning Technologies* 10, 2: 247–259. https://doi.org/10.1109/TLT.2016.2565476
9. Xi Chen, Paul N Bennett, Kevyn Collins-thompson, and Eric Horvitz. 2013. Pairwise Ranking Aggregation in a Crowdsourced Settingwsdm2013-preference-chen-et-al.pdf.
10. Kwangsu Cho and Charles MacArthur. 2010. Student revision with peer and expert reviewing. *Learning and Instruction* 20, 4: 328–338. https://doi.org/10.1016/J.LEARNINSTRUC.2009.08.006
11. Kwangsu Cho and Christian D. Schunn. 2007. Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers & Education* 48, 3: 409–426. https://doi.org/10.1016/J.COMPEDU.2005.02.004
12. Kwangsu Cho, Christian D. Schunn, and Davida Charney. 2006. Commenting on Writing: Typology and Perceived Helpfulness of Comments from Novice Peer Reviewers and Subject Matter Experts. *Written Communication* 23, 3: 260–294. https://doi.org/10.1177/0741088306289261
13. Kwangsu Cho, Christian D. Schunn, and Roy W. Wilson. 2006. Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology* 98, 4: 891–901. https://doi.org/10.1037/0022-0663.98.4.891
14. Shamil Chollampatt, Duc Tam Hoang, Hwee Tou Ng, and Hwee Tou Ng. 2016. Adapting Grammatical Error Correction Based on the Native Language of Writers with Neural Network Joint Models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1901–1911. Retrieved October 19, 2018 from http://www.aclweb.org/anthology/D16-1195
15. Shamil Chollampatt and Hwee Tou Ng. 2018. A Multilayer Convolutional Encoder-Decoder Neural Network for Grammatical Error Correction. *arXiv preprint* arXiv:1801, 08831. Retrieved October 19, 2018 from http://arxiv.org/abs/1801.08831
16. Shamil Chollampatt, Kaveh Taghipour, and Hwee Tou Ng. 2016. Neural Network Translation Models for Grammatical Error Correction. *arXiv preprint* arXiv:1606, 00189. Retrieved October 19, 2018 from http://arxiv.org/abs/1606.00189
17. Scott A. Crossley and Danielle S. McNamara. 2012. Predicting second language writing proficiency: the roles of cohesion and linguistic sophistication. *Journal of Research in Reading* 35, 2: 115–135. https://doi.org/10.1111/j.1467-9817.2010.01449.x
18. Scott A. Crossley, Tom Salsbury, and Danielle S. McNamara. 2012. Predicting the proficiency level of language learners using lexical indices. *Language Testing* 29, 2: 243–263. https://doi.org/10.1177/0265532211419331
19. Ronan Cummins and Marek Rei. 2018. Neural Multi-task Learning in Automated Assessment. *arXiv preprint* arXiv:1801, 06830. Retrieved October 17, 2018 from http://arxiv.org/abs/1801.06830
20. Nicola Dell and Neha Kumar. 2016. The Ins and Outs of HCI for Development. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, 2220–2232. https://doi.org/10.1145/2858036.2858081
21. Anders Ericsson, Ralf Krampe, and Clemens Tesch-Römer. 1993. The role of deliberate practice in the acquisition of expert performance. *Psychological review* 10, 1: 363. Retrieved April 16, 2019 from https://psycnet.apa.org/buy/1993-40718-001
22. KA Ericsson, RT Krampe, C Tesch-Römer - Psychological review, and undefined 1993. The role of deliberate practice in the acquisition of expert performance. *doi.apa.org*. Retrieved February 11, 2019 from http://doi.apa.org/journals/rev/100/3/363.html
23. Ilya M Goldin. 2012. Accounting for Peer Reviewer Bias with Bayesian Models. In *Proceedings of the Workshop on Intelligent Support for Learning Groups at the 11th International Conference on Intelligent Tutoring Systems.* Retrieved February 22,

2018 from https://pdfs.semanticscholar.org/d5a2/8c14e86afbc6f21a0bcc462404a529991cfe.pdf

24. Ilya M Goldin. 2012. Accounting for Peer Reviewer Bias with Bayesian Models. *Proceedings of the Workshop on Intelligent Support for Learning Groups at the 11th International Conference on Intelligent Tutoring Systems.*
25. Catherine M. Hicks, Vineet Pandey, C. Ailie Fraser, and Scott Klemmer. 2016. Framing Feedback: Choosing Review Environment Features that Support High Quality Peer Assessment. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, 458–469. https://doi.org/10.1145/2858036.2858195
26. Kevin G. Jamieson and Robert D. Nowak. 2011. Active Ranking using Pairwise Comparisons. 1: 1–9. https://doi.org/10.1523/JNEUROSCI.5514-03.2004
27. Rhelda Krugel and Elsa Fourie. 2014. Concerns for the Language Skills of South African Learners and Their Teachers. *International Journal of Educational Sciences* 7, 1: 219–228. https://doi.org/10.1080/09751122.2014.11890184
28. James A. Kulik and Chen-Lin C. Kulik. 1988. Timing of Feedback and Verbal Learning. *Review of Educational Research* 58, 1: 79–97. https://doi.org/10.3102/00346543058001079
29. Chinmay E. Kulkarni, Michael S. Bernstein, and Scott R. Klemmer. 2015. PeerStudio: Rapid Peer Feedback Emphasizes Revision and Improves Performance. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale - L@S '15*, 75–84. https://doi.org/10.1145/2724660.2724670
30. Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, and Scott R. Klemmer. 2015. Peer and Self Assessment in Massive Online Classes. *Design Thinking Research. Understanding Innovation.*: 131–168. https://doi.org/10.1007/978-3-319-06823-7_9
31. Diane Litman. 2016. Natural Language Processing for Enhancing Teaching and Learning. In *Proceedings of the 30th Conference on Artificial Intelligence (AAAI 2016)*, 4170–4176.
32. Kristi Lundstrom and Wendy Baker. 2009. To give is better than to receive: The benefits of peer review to the reviewer's own writing. *Journal of Second Language Writing* 18, 1: 30–43. https://doi.org/10.1016/J.JSLW.2008.06.002
33. Wencan Luo. 2017. Automatic Summarization for Student Reflective Responses. University of Pittsburgh. Retrieved January 15, 2018 from http://d-scholarship.pitt.edu/31455/
34. Gary Marsden, Andrew Maunder, and Munier Parker. 2008. People are people, but technology is not technology. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* 366, 1881: 3795–804. https://doi.org/10.1098/rsta.2008.0119
35. Andrew Maunder, Gary Marsden, Dominic Gruijters, and Edwin Blake. 2007. Designing interactive systems for the developing world - reflections on user-centred design. In *2007 International Conference on Information and Communication Technologies and Development*, 1–8. https://doi.org/10.1109/ICTD.2007.4937419
36. Chao Mbogo, Edwin Blake, and Hussein Suleman. 2016. Design and Use of Static Scaffolding Techniques to Support Java Programming on a Mobile Phone. In *Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education - ITiCSE '16*, 314–319. https://doi.org/10.1145/2899415.2899456
37. Fei Mi and Dit-Yan Yeung. 2015. Probabilistic Graphical Models for Boosting Cardinal and Ordinal Peer Grading in MOOCs. *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI-15)*: 454–460.
38. M. Molapo, M. Densmore, and L. Morie. 2016. Designing with community health workers: Enabling productive participation through exploration. In *ACM International Conference Proceeding Series*. https://doi.org/10.1145/2998581.2998589
39. M. Molapo, M. Densmore, and B. De Renzi. 2017. Video consumption patterns for first time smartphone users: Community health workers in Lesotho. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*.
40. Maletsabisa Molapo. 2015. Local Games, Local Content, Local Applications - On Mobile: ICTs for Informal Learning in Rural Africa. In *Digitally Connected: Global Perspectives on Youth and Digital Media*, Sandra Cortesi and Urs Gasser (eds.). Harvard Univeristy Berkman Center for Internet & Society, 54–58.
41. Maletsabisa Molapo and Melissa Densmore. 2015. How To Choose a Mobile Phone for an ICT4D Project. In *Proceedings of the Seventh International Conference on Information and Communication Technologies and Development Notes - ICTD '15*, 10–13.
42. Huy Nguyen, Wenting Xiong, and Diane Litman. 2017. Iterative Design and Classroom Evaluation of Automated Formative Feedback for Improving Peer Feedback Localization. *International Journal of Artificial Intelligence in Education* 27, 3: 582–622. https://doi.org/10.1007/s40593-016-0136-6
43. Benjamin D. Nye. 2015. Intelligent Tutoring Systems by and for the Developing World: A Review of Trends and Approaches for Educational Technology in a Global Context. *International Journal of Artificial Intelligence in Education* 25, 2:

177–203. https://doi.org/10.1007/s40593-014-0028-6

44. Chris Piech, Jonathan Huang, Zhenghao Chen, Chuong Do, Andrew Ng, and Daphne Koller. 2013. Tuned Models of Peer Assessment in MOOCs. https://doi.org/http://dx.doi.org/10.1002/ajh.24471
45. Chris Piech, Jonathan Huang, Zhenghao Chen, Chuong Do, Andrew Ng, and Daphne Koller. 2013. Tuned Models of Peer Assessment in MOOCs. Retrieved February 22, 2018 from http://arxiv.org/abs/1307.2579
46. Karthik Raman and Thorsten Joachims. 2014. Methods for ordinal peer grading. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*, 1037–1046. https://doi.org/10.1145/2623330.2623654
47. Karthik Raman and Thorsten Joachims. 2014. Methods for Ordinal Peer Grading. https://doi.org/10.1145/2623330.2623654
48. Marek Rei and Helen Yannakoudakis. 2016. Compositional Sequence Labeling Models for Error Detection in Learner Writing. *arXiv preprint* arXiv:1607, 06153. https://doi.org/10.18653/v1/P16-1112
49. Thomas Staubitz, Dominic Petrick, Matthias Bauer, Jan Renz, and Christoph Meinel. 2016. Improving the Peer Assessment Experience on MOOC Platforms. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale - L@S '16*, 389–398. https://doi.org/10.1145/2876034.2876043
50. Kaveh Taghipour and Hwee Tou Ng. 2016. A Neural Approach to Automated Essay Scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1882–1891. Retrieved October 17, 2018 from https://www.kaggle.com/c/asap-aes
51. Fatema Wali and Henk Huijser. 2018. Write to improve: exploring the impact of an online feedback tool on Bahraini learners of English. *Learning and Teaching in Higher Education: Gulf Perspectives* 15, 1. https://doi.org/10.18538/lthe.v15.n1.293
52. Andrew E. Waters, David Tinapple, and Richard G. Baraniuk. 2015. BayesRank: A Bayesian Approach to Ranked Peer Grading. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale - L@S '15*, 177–183. https://doi.org/10.1145/2724660.2724672
53. Andrew E. Waters, David Tinapple, and Richard G. Baraniuk. 2015. BayesRank: A Bayesian Approach to Ranked Peer Grading. *Proceedings of the Second (2015) ACM Conference on Learning @ Scale - L@S '15*: 177–183. https://doi.org/10.1145/2724660.2724672
54. Wenting Xiong. 2014. Helfulness-Guided Review Summarization. University of Pittsburgh. Retrieved January 15, 2018 from https://d-scholarship.pitt.edu/22509/1/WentingXiong_ETD_12042014_final.pdf
55. Helen Yannakoudakis, Øistein E Andersen, Ardeshir Geranpayeh, Ted Briscoe, and Diane Nicholls. 2018. Developing an automated writing placement system for ESL learners. *Applied Measurement in Education* 31, 3: 251–267. https://doi.org/10.1080/08957347.2018.1464447