Keystrokes, Edit Distance, and Grading Rules: Psychometric Properties of Short Answer Items

David Lang
Stanford University
davidnathanlang@stanford.edu

Ben Stenhaug
Stanford University
stenhaug@stanford.edu

Rene Kizilcec
Cornell University
kizilcec@cornell.edu

ABSTRACT

This research evaluates the psychometric properties of short-answer response items under a variety of grading rules in the context of a mobile learning platform in Africa. This work has three main findings. First, we introduce the concept of a differential device function (DDF), a type of differential item function that stems from the device a student uses to take an assessment. Second, we identify a plausible mechanism for this DDF by examining the keystroke requirements of smartphone and basic mobile phone users. We identify a set of platform design rules to mitigates this bias. Lastly, we suggest that the edit distance of student responses can be used as a tuning parameter to optimize the Cronbach's alpha of the assessment. We find that literal string evaluation performs poorly compared to other grading rules. Partial string matching with an edit distance of two provides the highest reliability across exams. This is a simple yet effective rule, which performs well across a variety of assessments.

KEYWORDS

Mobile Learning, Natural Language Processing, Psychometrics, Grading

ACM Reference Format:
David Lang, Ben Stenhaug, and Rene Kizilcec. 2019. Keystrokes, Edit Distance, and Grading Rules: Psychometric Properties of Short Answer Items. In *Proceedings of American Educational Research Association (AERA 2019).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

The design and form of educational instruction are often limited by the medium and the nature of the interaction. Educational platforms often have the choice between having extensive features or developing parsimonious tools that can work on a variety of devices. This fact is most evident in educational initiatives in developing countries where resource constraints are paramount. Transmission of educational content via textbooks, direct instruction, or even online platforms are often impractical or too costly. As such, there is a growing effort to leverage existing technological resources and assets for educational purposes. Cellular phones have been of particular interest because their adoption has been rapid. Estimates of global cell phone ownership now exceeds 60% and the number is likely to continue to grow rapidly [**?**]. Moreover, cellphones are a growing part of non-governmental organizations' attempts to improve living standards not only by increasing communication infrastructure but also by using these devices to provide access to health services and financial markets [1–3]. In addition to being accessible to a large population, the advantage of learning via mobile devices means that interventions can benefit geographically isolated individuals, as well as individuals who are unable to attend schools due to socio-cultural barriers.

The fact that educational content is being designed for cell phones also informs the design choices of a platform. The first design choice is the channel of communication: internet, multimedia message service, or short message service (SMS). If a platform decides to make their content available through either the internet or MMS, this yields a richer platform with the potential for communicating with both images and text but effectively limits the pool of potential users. For instance in Kenya, smartphone ownership is only 26% but overall ownership of cellphones is 82% [**?**]. Moreover, reliable connections to the internet are not necessarily available to all parties. The most extensive means of communication for cellular phones is SMS but it constrains a platform to communicate exclusively through text. SMS messages allow individuals to text and send 160 character messages to one another. The fact that many cellphone carriers still charge users on a per-message basis prompts the need for judicious use of messages. This informs the design of assessments and evaluation on these platforms. For instance, writing high-quality multiple-choice items often means that each distractor could result in meaningful financial costs to students.

We evaluate short-response data from a text-message-based learning platform called Eneza. Eneza is a learning platform that operates in Kenya and offers multiple services with mobile devices. The service offers lessons and short-answer-based assessments in a variety of subjects. Rather than provide multiple-choice items with lengthy distractors, the service provides short-answer and fill-in-the-blank style prompts that utilize keyword matching. The service's current grading algorithm detects whether or not the exact keyphrase is contained in a student's response. This paper explores the psychometric properties of the service's items under a variety of grading rules. In particular, given the nature of the platform, typographic errors (typos) are likely to occur and modest error correction may improve the performance of these items [4].
We also focus on how using a smartphone versus a basic mobile device could influence performance on these items and learning with the platform. If items attempted on a smartphone are easier due to the ergonomics of the device e.g.(access to a full keyboard, automatic spellcheck, faster data access, etc.), then there are implications for both the design of the platform and utilization of this type of data. Namely, if this type of platform were to be used

AERA 2019, April 2019, Toronto ,Canada

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

for evaluative purposes, whether or not a user was responding via a smartphone could bias estimates of an individual's ability. We explore the potential for this type of differential item function by examining the number of keystrokes required on a smartphone versus a basic mobile device.

2 LITERATURE REVIEW

2.1 Automated Grading

Previous research on automated grading of short-answered responses has historically focused on internal validity of grading systems compared to a hand-coded evaluation of responses. Typically these grading systems use a host of string matching and natural language processing techniques to identify statements that students should include in their response. This work has typically found that on such measures such as inter-rater agreement or Cohen's Kappa, these systems perform reasonably well [5]. A key limitation to this body of work is that while these measurements can inform the relative agreement of human versus machine-grading, this work fails to address to the validity of the underlying exam. There exists a rich literature on assessing the validity and reliability of exams and surveys. The most common measure used is Cronbach's alpha [6], Cronbach's alpha is a measure of internal consistency of sets of items. It is calculated as follows:

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K} \sigma_{Y_i}^2}{\sigma_X^2}\right)$$

σ_x^2 corresponds the variance of the overall test scores. $\sigma_{Y_i}^2$ corresponds to the variance of item i , and K corresponds to the number of items. Intuitively, this measure can be thought of as the associated correlation between split-halves of the same exam. The benefit of machine grading is that alternative scoring rubrics can be quickly evaluated by these sorts of measures whereas human grading can take extensive time. Rather than choosing a grading rule that maximizes a form of human-interrater agreement on a single item, we propose an alternative metric that chooses a grading rule such that Cronbach's alpha is maximized across the entire exam. This approach is a common machine learning framework.

Another common critique against machine grading of essays and short answer questions is that humans are generally better graders than their machine counterparts. Critics contend that they reward students for being verbose or adding in irrelevant text [7]. We would like to give pushback on this notion of human superiority and present a case where human grading would be substantially inferior to machine grading. Human graders have been shown to exhibit much more idiosyncratic behaviors and are subject to other form of cognitive biases. For instance, human graders are unable to consistently implement grading policies across a variety of instances. When asked to categorize words by concept, individuals were less likely to classify homophones consistently (vain,vein, or vane) [8]. A human designed answer-key may be prone to some of these biases.

In our dataset, these types of issues are further amplified in that we are dealing with student input in a variety of languages on a variety of devices. The nature of these devices in turn inform how students will input their answers to test questions. For instance, students using a smart device will input their responses with a full or virtual keyboard. These students are likely to commit relatively modest typos. These students typically have to enter one character of input to produce one character on the screen.

Students utilizing a phone without a full keyboard are more likely to use a multitap or T9 interface. Both of these texting techniques rely on the fact that each digit in a basic mobile phone maps to a set of letters in Table 1. Multitap input requires users to hit each key repeatedly to cycle through a set of letters associated with each key. For instance, to enter a 'b', a user will have to enter the '2' key twice. Multitap users thus will have a very different set of key entries to produce the desired input as a virtual keyboard user. Potentially, multitap users will have to enter three times as many keys for the equivalent length text on a virtual keyboard. An alternative to multitap is T9 input mapping. T9 mapping allows an individual to enter one key per character entered. At the end of each word, a user is then prompted with a choice of words associated with that input string. For instance, entering the numbers 4663 on a T9 text messaging system map to a number of common words:good,gone,home,hone,.... These 'textonyms' are more likely to be produced by students who are using a mobile device [9]. Due to the nature of these input methods, identical number sequences are likely to produce very dissimilar words based on traditional methods of edit distance.

Traditionally, edit distance is measured as the number of character insertions, deletions , substitutions, and transpositions required to transform one string into another [10]. To convert the string 'parse' to 'page', the minimum edit distance of this transformation is two (one substitution and one deletion). For instance the words 'equitable' and 'fruitcake' have an identical T9 entry string but have a Levenshtein edit distance of 5.

2.2 Ergonomic Considerations

These design considerations are quite important because they have very clear and direct impacts on a student's ability to process and engage with material. Expert T9 users can typically type at double the rate of expert multitap users [**?**]. Other considerations suggest that non-smart devices generally have lower performance quality in terms of both speed and accuracy. Past research has suggested that multitap is the slowest form of messaging [11]. To the extent that there is a smartphone gap across socioeconomic status, texting features could exacerbate these gaps because students with basic phones are unable to benefit from the efficient typing speed.

More generally, there's been substantial work on how best to design predictive text and error correction systems. Experiments that artificially induced typing errors found that the likelihood of individuals catching and revising typos drops off precipitously for longer words [12] . This work suggests that reducing keystrokes may improve student accuracy. With respect to autocompletion, research has found that these features tend to reduce the number of user keystrokes but also tend to reduce a user's speed [13]. As such, it's not entirely clear whether these features should be incorporated into educational technology.

3 DATA

The dataset we are using is from an SMS text message service called Eneza. In total, we utilize data from 499,796 responses from a set of

Table 1: Nine-Key Mapping

Phone Key	Letter
2	abc
3	def
4	ghi
5	jkl
6	mno
7	pqrs
8	tuv
9	wxyz

open-response items. These responses were generated from 33,198 students. The exams were generated by 135 unique content creators . These responses were gathered from 490 exams and 2,443 items. These questions utilized literal string matching and checked if the terms matched a set of keywords. A sample problem is below in table 2. In this instance, a student was asked to list two examples of insects after reading a passage about the subject. The student entered his response and it was matched against a set of words within the key. There are several notable features about this example problem. First, the answer key uses an inconsistent ruling of whether the response should be singular or plural. For instance, the answer key contained both the terms 'fly' and 'flies' but only had the plural form 'bees'. In this instance, the student only received half-credit because they used the singular form of butterfly. The other interesting feature about this response and many like it is that the system grades based on exact string matching rather than partial string matching. For instance, if the student in this example typed 'honeybees' instead of 'honey bees', the student would have gotten zero credit for the item.

3.1 Dictionary T9

We tokenized each of the responses in the problem answer keys. In total, there were 5,142 distinct terms as part of these answer keys. We then matched these terms against a set of preexisting dictionaries of words that are commonly included in T9 libraries. We found that approximately 53% of these unigrams had a corresponding entry in a T9 dictionary.

We then computed the number of keystrokes for each potential answer under several different scenarios. In one scenario, we assume that users type exclusively using multitap keystrokes. In the other scenario, users type exclusively using T9 keystrokes. In the event that the term does not exist within the T9 dictionary, the user then has to reenter the whole term using multitap. In both cases, we assume that each student types perfectly and is simply unaware whether or not their term is in the T9 dictionary. Figures 1 and 2 display the relative efficacy under both strategies. If a student is entering a word that is in the T9 dictionary, it requires exactly as many keystrokes as the number of letters in the term. If the term is not present, utilizing T9 input will resort in unnecessary typing for an individual.

Based on this information, we then explored the possibility of whether there's substantial heterogeneity by type of question or quiz. Identifying whether dictionary terms or non-dictionary terms

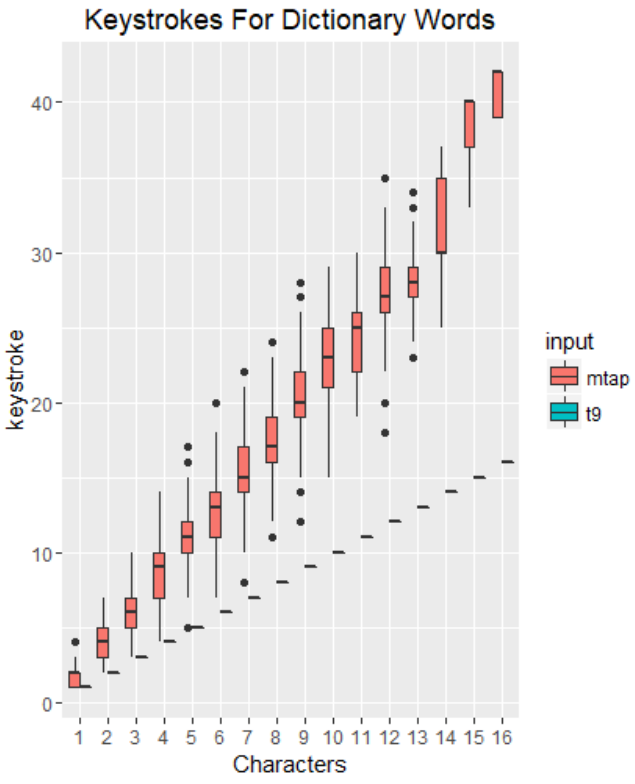

Figure 1: Keystrokes for Dictionary Words

are the predominant form of answers could inform how students are instructed to attempt these items. We found that the overwhelming majority of quizzes were heavily concentrated with T9 terms. Figure 3 shows that the median quiz and question have approximately 80% of their associated terms in a T9 dictionary. Those terms that were not considered part of the T9 dictionary tended to be concentrated in either numerical calculations, proper nouns, foreign terms, or alternative spelling variants e.g.(theater versus theatre). One natural implication from this finding is that users should be encouraged to use T9 communication as a default unless they are interacting with a quiz that has a particularly low concentration of dictionary terms. In fact, if users are typing with a strategy that minimizes keystrokes, we found that the optimal strategy saves approximately one standard deviation of keystrokes or roughly four characters per word. The associated t-test is statistically significant with t-statistic of 59.8. [1]

4 GRADING

As we encountered in our earlier example, Both students and content creators exhibit substantial idiosyncractic behavior in terms of how they express answers. Students may not fully understand the specificity required to match the answer key. Content creators may not consistently write answer keys such that they use the same form of singular/plural agreement or spelling conventions.

[1] The associated degrees of freedoms is based on 2,173 questions. Numeric questions were excluded from this analysis.

Table 2: Example Problem and Response

Feature	Example
Question	List the examples of organisms under class insecta.
Answer	Examples of organisms in class insecta;dragon fly,weevils,beetles,termites,locusts,blowflies,silk worm bees,butterflies,tsetse fly,mosquitoes,flies.moths.
Student Response	butterfly and honey bees
Key	dragon, fly, weevils, beetles, termites, locusts, blowflies, silk, worm, bees,butterflies, tsetse, fly, mosquitoes, flies, moths butterflies, tsetse, fly, mosquitoes, flies, moths

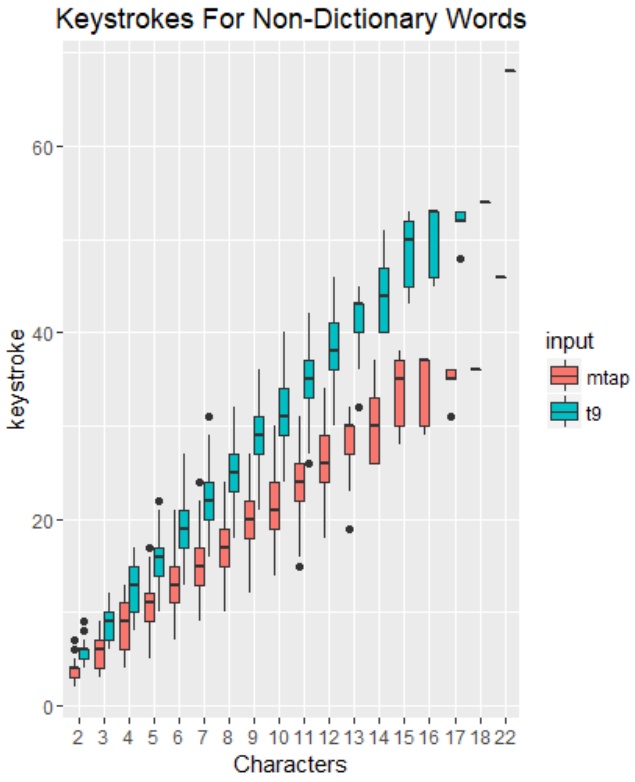

Figure 2: Keystrokes for Non-Dictionary Words

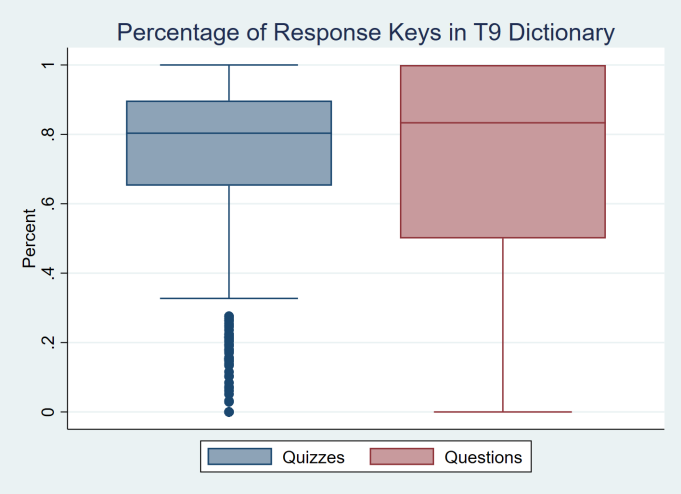

Figure 3: Concentration of T9 Terms By Quiz and Question

As such we explore how quizzes change under alternative grading schemes. We briefly considered grading policies based on stemming and lemmatization of answers but dismissed them as these transformations would not consistently address misspellings, alternative variants. or proper names. Additionally, these resources require a preexisting library of terms. Edit based distance measures can be generated as long as there is an answer key.

As such, we began by regrading items based on alternative scoring rules based on the edit distance between the student's submission and the answer key. We allowed zero to four degrees of edit distance and graded each response based on this criterion. One clear results of this transformation can be seen in Figure 4 . The light orange shading corresponds to items graded with an edit distance of zero. Under this grading schema, the modal item correctness is zero and a third of items have less than ten percent correctness. When items are graded within one or more degree of edit distance, these items follow a much more normal distribution and likely could be used for evaluative purposes. Also, the fact that these edit distances rules seem to be capturing misspellings and typos may be key to maintaining student engagement. Students may experience more frustration and leave a platform if they are marked incorrectly due to a typo or other trivial error.

Increasing the edit distance of a grading rule mechanically increases item correctness. However as mentioned earlier, there are other measures for calculating and measuring the validity of assessments. We compute Cronbach's alpha across these five grading schemes for each quiz[2]. The results can be seen in Figure 5. This figure indicates that increasing the edit distance by one or two character entries tends to improve the reliability. Going from an edit distance of zero to an edit distance of one results in a 5.88% percentage gain in the quizzes reliability. We then tested a one-way ANOVA of these edit distances on reliability. We rejected equivalence of reliability across groups with an associated F-statistic of

[2]In the event that Cronbach's alpha cannot be calculated due to rank deficiency, we assume a alpha to have a value of -1. We chose this value because it is the most severe penalty that is still a feasible value for alpha. This finding is robust to other possible imputed values for Cronbach's alpha

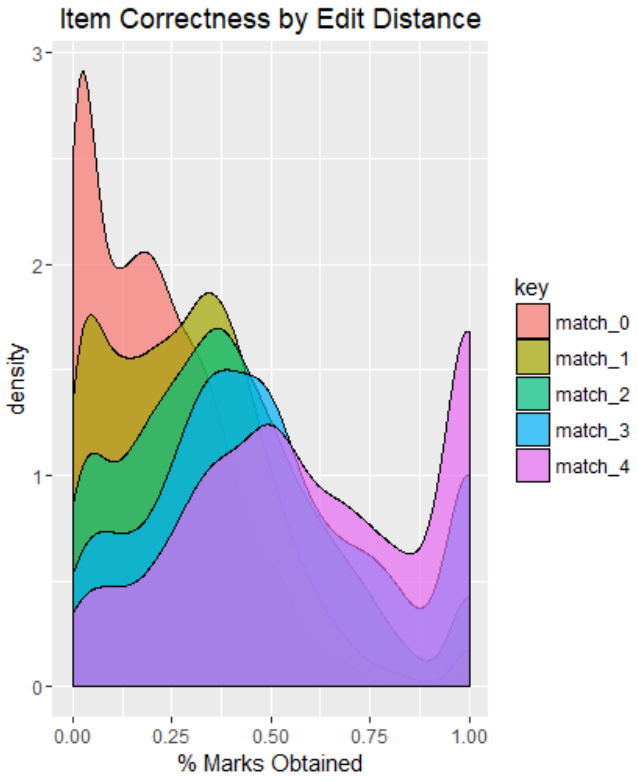

Figure 4: Item Correctness by Edit Distance

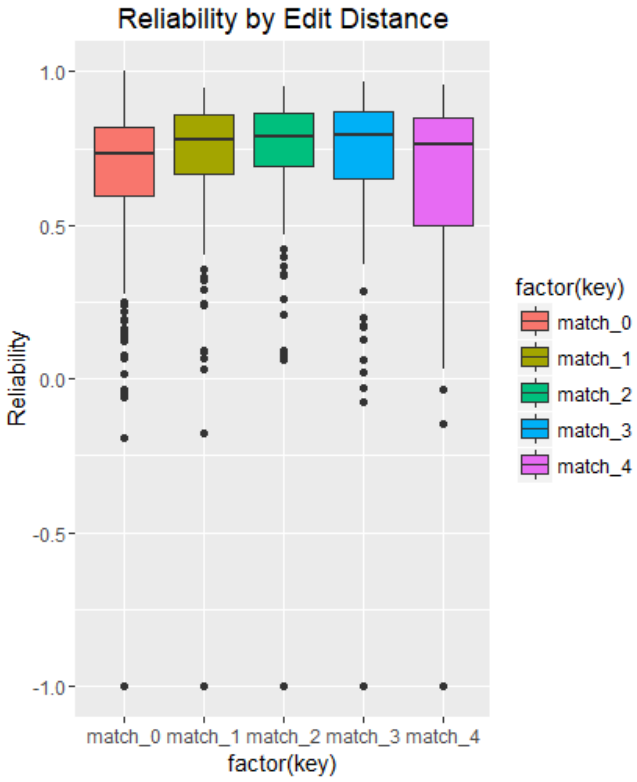

Figure 5: Cronbach's Alpha by Edit Distance

15.33 . These findings suggest that mild increases in edit distance may in fact improve a quiz's reliability.

5 LIMITATIONS TO ANALYSIS

Ultimately, good test design depends on multiple measures of validity and reliability. While increasing edit distance seems to improve item correctness and Cronbach's Alpha, the underlying construct of what is being measured could be changing as a result of these new grading methods. For instance, many terms in the sciences have their meaning changed by a single character or two e.g.(abiotic and biotic, exothermic and endothermic, etc.) Allowing these terms to be treated as equivalent could impede students' learning. Moreover, the notion of reliability may not be the optimal goal of an item. Items designed to teach rather than assess students may have different performance requirements.

6 CONCLUSIONS

In this paper, we have documented several key features in designing and implementing learning technology tools in an SMS platform. We found that encouraging students to use T9 texting substantially reduces a student's keystrokes. This information suggests that slight changes and prompting could improve students' ability to consume content. We also found that student responses are highly sensitive to slightly different grading policies. This suggests that multiple scoring forms should be utilized for evaluating and grading items. Future work will attempt to gain better knowledge of users. We hope to explore the extent to which these forms of misspellings and typos can advance our understanding of achievement gaps.

The other goal of future work is to generate algorithms that are more cognizant of misclassification and misgrading. Sensitivity analyses that focus on the proportion of characters that require alteration rather than the number of edits may prove more robust and useful to other grading contexts . Additionally focusing on subject areas and their respective sensitivities to these grading rules may inform how content could be better designed. Finally, We have presented two pieces of work on using an alternative repesentation of words (multitap and T9). Existing NLP technology can represent words in dense embedded spaces and can represent words that have similar meanings but very different or distinct character representation. Identifying whether character or word embeddings similarity measures can be used for grading student responses is a rich area for future study [14] [15].

ACKNOWLEDGMENTS

The authors would like acknowledge the Institute for Educational Sciences and Grant Number R305B14009.

REFERENCES

[1] RT Lester, P Ritvo, EJ Mills, A Kariri, and S Karanja. Effects of a mobile phone short message service on antiretroviral treatment adherence in Kenya (WelTel Kenya1): a randomised trial. *The Lancet*, 2010.
[2] JC Aker and IM Mbiti. Mobile phones and economic development in Africa. *The Journal of Economic Perspectives*, 2010.

[3] N Hughes and S Lonie. M-PESA: mobile money for the âĂIJunbankedâĂİ turning cellphones into 24-hour tellers in Kenya. *Innovations*, 2007.
[4] Bo Han and Timothy Baldwin. Lexical Normalisation of Short Text Messages: Makn Sens a #twitter. pages 368–378, 2011.
[5] Steven Burrows, Iryna Gurevych, and Benno Stein. The Eras and Trends of Automatic Short Answer Grading. *International Journal of Artificial Intelligence in Education*, 25(1):60–117, 3 2015.
[6] Lee J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334, 9 1951.
[7] KE Lochbaum, M Rosenstein, and P Foltz. Detection of gaming in automated scoring of essays with the IEA. *National Council on*, 2013.
[8] Veronika Coltheart, Karalyn Patterson, and Judi Leahy. When a ROWS is a ROSE: Phonological effects in written word comprehension. *The Quarterly Journal of Experimental Psychology Section A*, 47(4):917–955, 11 1994.
[9] M Kamvar. *Using context to improve query formulation and entry from mobile phones*. 2008.
[10] VI Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 1966.
[11] L Butts and A Cockburn. An evaluation of mobile phone text input methods. *Australian Computer Science Communications*, 2002.
[12] AS Arif and W Stuerzlinger. Predicting the cost of error correction in character-base... - Google Scholar. *Proceedings of The SigCHI COnference*, 2010.
[13] Philip Quinn and Shumin Zhai. A Cost-Benefit Study of Text Entry Suggestion Interaction. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, pages 83–88, New York, New York, USA, 2016. ACM Press.
[14] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 12 2017.
[15] X Zhang, J Zhao, Y LeCun Advances in neural information, and undefined 2015. Character-level convolutional networks for text classification. *papers.nips.cc*.