Developing items for a 1st grade screening test in reading: testing some basic principles

Arild Michel Bakken, arild.m.bakken@uis.no, Associate Professor, Norwegian Reading Centre

Per Henning Uppstad, per.h.uppstad@uis.no, Professor, Norwegian Reading Centre

Bente Rigmor Walgermo, bente.r.walgermo@uis.no, Associate Professor, Norwegian Reading Centre

Keywords: screening tests, multiple-choice items, reading difficulties

Abstract

A strong body of evidence has emerged supporting the validity of multiple-choice items in different contexts and providing insights into good practice when it comes to developing and validating items (See e.g. Haladyna, 2004). However, both item content and item design must be adapted to the purpose of the testing. General guidelines derived from theory on mainstream reading development apply optimally to tests for normally distributed populations, as they focus on how the full range of a skill can be measured. For tests with a more specific purpose, such as identifying those who are at risk for reading and writing difficulties, it may be advisable to violate some generally recognized guidelines and to formulate new ones.

In this study we take the first steps towards formulating tentative guidelines on the design of multiple-choice items when the purpose is to identify struggling readers through group administered digital tests. The empirical grounds for the study come from three sources: a) A list of criteria describing our thinking during item development for a large-scale screening test is the deductive part of our data material. We confront these theoretical assumptions with two empirical sources: b) think-alouds with at-risk pupils during informal testing and c) a quantitative data material, namely the statistics from a large pilot study of the test items.

The approach taken emphasizes the application of knowledge on the unique considerations of those students who are likely to fall behind. The assumption is that this systematic focus will strengthen the predictive validity if not also concurrent validity. As such, the present study will contribute to our knowledge some tentative guidelines for item development for screening tests in reading.

Introduction

A strong body of evidence has emerged supporting the validity of multiple-choice items in different contexts and providing insights into good practice when it comes to developing and validating items (See e.g. Haladyna, 2004). Necessarily, however, both item content and item design must be adapted to the specific purpose of the test. For tests with the specific purpose of identifying children who are at risk for reading and writing difficulties, general principles of item development must be filled with specific content and possibly be adapted. They must also be complemented by subject-matter specific principles related to the construct being

tested. The level of mastery of reading skills at end of first grade differs across orthographies. Even more so, it differs what reading skills low performing readers master. In order to identify those students who are likely to fall behind, it is necessary to estimate the level of difficulty from where to start when creating test items. To put it short, we need to track the orthographic difficulties that are specific to the struggling readers, and create items from this point.

In this study, we take some initial steps towards formulating tentative principles on the design of multiple-choice items when the purpose is to identify struggling readers through group administered digital tests.

Our study was carried out in the context of the development of new screening tests in reading for 1st graders in the Norwegian school system (six-year-olds). The test was to consist of 20 items of word reading and 20 items of sentence reading. The latent construct behind these items were decoding and reading comprehension respectively (cf the "simple" view of reading (Gough & Tunmer 1986)). All items were multiple choice items. In the early stages of test development, we organized a large-scale pilot of the items developed thus far, and that pilot is the empirical basis for this study, together with a think-aloud we will do with at-risk children solving some of the items.

The fact that cut-offs in screening tests for reading are set by convenience levels (Kane, 2017), is partly based on the experience that no clear qualitative difference in performance can be found among students at the e.g. 10, 15, 20, 25, 30 percentile. However, despite the continuity of performance, we hypothesize that there are some errors that very poor readers are more likely to make than readers without a risk of developing reading difficulties. In the existing framework for the Norwegian screening tests in reading, one requirement for test development is that items should be mastered by between 70 and 90 % of the students. This means that if an item was shown to be mastered by only 65 % of the students, it would have been discarded as too difficult, and if mastered by 95% as being to easy. While this is only one requirement among several in a sophisticated framework, importantly, it applies as a first step in test development, sorting plausible item candidates from not plausible ones. The potential problem in this, is the discarding of very simple items that despite their ease may serve as first row identifiers for the targeted students. Also, the fact that the easiest items have been discarded by the requirements stated in the existing framework, makes it plausible that such easy items are avoided in the process of item development, with a lack of knowledge on how to build very easy items of high quality as a consequence. As will be shown below, a number of the items mastered by more than 90 % of the students show high psychometric quality. As a consequence, one may inadvertently have come to dilute the power of tier 1 items, in favor of tier 2 items - a item tier that is likely to elicit errors typical for both general development and reading difficulties. If so, the outcome is likely to be a test with sub-optimal longitudinal prediction.

In the exploration that follows, we will apply a heuristic model of how to approach the issue of item development for this specific purpose. In this, we consider it convenient to focus on two sets of items following somewhat different principles or criteria. The inner tier represents items where more than 90% of the children succeed. These kinds of items would normally be

discarded from the final test due to the specifications in the framework considering the whole population. The outer tier represents items where 70%-90% of the children succeed, in conformity with the framework of the screening test. We hypothesized that different kinds of principles would apply for these two tiers.

Several kinds of principles could be relevant for a screening test such as this, both principles stemming from general test theory and principles stemming from reading theory. The field of general test theory has produced many principles or guidelines that are fairly well established and consensual, and which all item developers would be well advised to observe. One such general principle is the guideline stating that test developers should "make all distractors plausible" (Haladyna, 2004: 120). In our context, this entailed that all distractors should be words resembling the right answer in some way, or situations having some things in common with the right answer. It also entailed that, for the word reading items, all distractors should be words existing in Norwegian language (not nonsense-words). However, knowing what might seem plausible to a six-year-old, is not self-evident, and the pilot results as well as the qualitative try-out allowed us to confront our assumptions regarding this to empirical data.

Another important general test theory guideline was the one stating that test developers should "assign the position of the right answer randomly" (Haladyna, 2004: 113). Although this guideline might be perfectly valid for tests on mainstream populations (with the caveat concerning "edge aversion" posed by Haladyna with reference to Attali and Bar-Hillel (2003)), we made an assumption that when the objective is to identify the 20% weakest readers it might be advisable to derogate from this guideline in some cases. Because of previous experience with test development for this age group, we suspected that when the options are words (such as in our word reading items), the weakest readers might prove, instead of edge averse, "end averse" – showing a preference for the first options presented. In order to achieve good discrimination at the lower end of the spectrum, it could therefore be advisable to distribute the right answers with a slight overweight towards the end.

A second pillar of item development principles comes from reading theory, and here the picture is much more unclear and subject to dispute. The objective is not only to identify the weakest readers in a concurrent perspective, but to predict who will struggle with reading at a later stage.

Such principles can be divided into a couple of broad categories. One category which discriminates highly between different levels of mastery concerns letter knowledge. The order in which different letters are known to children of course varies between orthographies. For our purposes, we were leaning on data from a previous screenings for letter knowledge in connection with the development of a computer game (Njå 2019), identifying to what extent the individual letters are known in a population. This resulted in a list of increasing difficulty for letters and complex graphemes based on multiple sources:

In a first level the order of the items i-r was set to be identical with the order of how letters were presented in three typical ABC books (mean values) and which are mastered by all students in a screening test.

i – l – s – o – e – a – m – r

In a second level the order of the items u-j was set to be identical with the degree of mastery (percent) in a population of students in a Norwegian screening test for first grade (< 100 %)

u – t – b – f – n – v – k – å – h – p – d – g – æ – y – ø – j

For complex graphemes and diphthongs, the order of the items was set to be identical with the order of mastery (percent) in a sample of dyslexic students (n=61):

ei – øy – au – ai – sj – ng – skj – kj – kj – gj – hj

It can be assumed that words including only letters from the first group will be extremely easy, whereas including letter from the second group will increase difficulty gradually, the further right one goes on that axis. Item creation, then, could start in the first level, gradually increasing the difficulty. Knowing that while doing so, the more complex it gets the more the items are likely to induce normally developing readers into making errors.

By starting out here, we start tracking at the edge of where at-risk students -i.e. the primary target students of this test - make errors that are unique to this group. In line with the reasoning of Walgermo, Bakken & Uppstad (In preparation), we argue that we need to start in this end, and gradually include items of larger complexity until a convenient cut-off is reached. In order to be able to do so with rigor, an account of increasing complexities in the actual orthography is required.

Distributing difficulty evenly on a scale is relatively easy then, when it comes to individual letter knowledge. For other principles, this is less straightforward, and therefore even more interesting to test empirically. We divide these other principles into three broad categories according to whether they tap difficulties related to the visual manifestation of the sign or its phonetic manifestation – or whether it is a mix between the two.

An example of a visual principle could be using distractors beginning and ending in the same way as the stem, tapping unclear word images in the beginning reader. Alternatively, the distractors could only begin in the same way, tapping a tendency not to read the word in full. The difficulty could also be on the level of the single grapheme. In particular, graphemes that are mirror images of each other, such as *b* and *d*, *b* and *p*, often cause difficulties.

Examples of phonetic principles could be exploiting phonemes that are close to each other such as /ɨ/ and /y/ or /y/ and /ʉ/ - or more generally any distinction relying only on a change of one vowel. When it comes to consonants, one could also exploit similarities between single phonemes, such as /p/ and /b/, or the complexity of consonant clusters.

Some types of difficulties are difficult to attribute mainly to either visual or phonetic problems. This includes phonemes that could be conveyed in several different graphemes (in Norwegian this is the case with /u/ - sometimes written *u*, sometimes *o*), phonemes that do not have a grapheme (such as /ŋ/). It also includes cases where the name of the grapheme is taken for the phoneme (for example *gav* instead of *gave* [=*gift*] – knowing that the grapheme *v* is pronounced /ve/), and it includes the permutation of letters such as choosing *risp* [=scratch] instead of *rips* [=*redcurrant*]. It includes, finally, over-orthophonic spelling of words that in Norwegian are not completely orthophonic.

All of these phenomena are known to cause difficulties for certain types of readers, and could therefore constitute subject matter specific principles for developing a screening test in reading. The problem is knowing exactly to whom, and thus where on the ability scale they discriminate. To item developers this is essential information.

Method

As mentioned, our study was carried out in the context of the development of new screening tests in reading for 1st graders in the Norwegian school system (6 years old). The test was to consist of 20 items of word reading and 20 items of sentence reading. In the word reading items, the stem was an image representing an object, and the options were four words, of which one corresponded to the object represented in the image. In the sentence reading items, the stem was a sentence representing a situation, and the options were four images of which one corresponded to the situation represented.

Leaning on the principles mentioned above, we developed around 200 items of word reading and 200 items of sentence reading. In the early stages of test development, we organized a large-scale pilot of these items. They were divided into 10 testlets of a presumably similar level of difficulty, each with ca. 20 items of word reading and 20 items of sentence reading. Each of the testlets were given to a sample of ca. 500 children towards the end of first grade. The testlets were analyzed for the purposes of further development using both classical analysis and IRT analysis for difficulty level and discrimination.

For the purposes of this study, we selected 4 testlets randomly for a preliminary qualitative analysis of the functioning of our principles. We first examined how well represented each of the four positions in the options were, in order to check for "edge aversion" or "end aversion".

In conformity with our two-tier approach we then put the items where more than 90% of the children succeeded (excluding missing) in one pile, and the items where 70%-90% succeeded in another. The few items where less than 70% succeeded were discarded. In the pile of items with more than 90% correct answers, we discarded items where no distractor had attracted more than 1% of the answers. In each of the two piles, we discarded items where discrimination was <0,85 and we took a closer look at the 8-10 items with the highest discrimination.

With the best performing items, we asked ourselves which principles we had used during item development and we compared to other items having used similar principles, both from the same pile or from the other pile, or even from the pile of discarded items. In this process, we asked ourselves questions like: "why was this item more difficult than this one?" or "why did this item discriminate better than this one?" We also checked what kind of principles were represented in each of the piles.

The next step of this study will be to subject a certain number of items to a more in-depth qualitative try-out primo September. We will subject four previously identified at-risk pupils to the chosen items, and do a think-aloud with them as they try to solve the items. The objective is to ascertain whether our analysis of the functioning of the items is accurate, and to gain new insight into the functioning of the items from the test taker's perspective and the cognitive processes going on in the struggling reader. The results from this try-out will be ready before the Baku conference.

Results

Edge aversion vs. end aversion.

A preliminary analysis of the four selected testlets (79 word reading items) in terms of the relationship between the number of times each position was the right answer and the number of times it was chosen, yielded the following:

	Testlet 1			Testlet 2			Testlet 3			Testlet 4			Total		
	Right	Selected	Diff.	Right	Selected	Diff.	Right	Selected	Diff.	Right	Selected	Diff.	Right	Selected	Diff.
Position 1	0,05	0,10	0,05	0,20	0,23	0,03	0,15	0,19	0,04	0,21	0,22	0,01	0,15	0,18	0,03
Position 2	0,15	0,11	-0,04	0,10	0,13	0,03	0,15	0,16	0,01	0,16	0,18	0,02	0,14	0,14	0,00
Position 3	0,30	0,28	-0,02	0,30	0,26	-0,04	0,35	0,30	-0,05	0,21	0,20	-0,01	0,29	0,26	-0,03
Position 4	0,50	0,42	-0,08	0,40	0,32	-0,08	0,35	0,30	-0,05	0,42	0,36	-0,06	0,42	0,35	-0,07

The analysis shows a strong tendency toward end aversion in this sample, i.e. the children prefer options situated to the left to those situated to the right. Option 1 is the most over-represented in the children's choices. Option 2 is only slightly over-represented, whereas option 3 and 4 are under-represented and option 4 heavily so. The overall analysis shows no tendency toward edge aversion (except in testlet 4, where option 2 is more over-represented than option 1, which could be a slight indication of edge aversion).

This analysis comforts our hypothesis that young children are susceptible to prefer options situated to the left, and that it would be advisable, when developing items for screening tests for this age group, to distribute the right answers with a slight overweight towards the end. However, in order to test this more thoroughly we would need to test it on items with evenly distributed answers. In this pilot, we integrated our hypothesis in the development and privileged position 3 and 4 quite heavily. This surely has an effect on the over-/under-representation of the positions. Even with this caveat in mind, testlet 4 for example, where position 1 is more selected than position 3, although they have the same number of right answers, shows the presence of end aversion.

Principles concerning the visual manifestation of the sign.

A preliminary qualitative analysis of the best performing items show some interesting differences between tier 2 and tier 1 items, i. e. items with a success rate of 70%-90% and items with a success rate of over 90%.

The best performing item in tier 2, with a discrimination of 1,7, was *dråpe*:

Option 1, *bråte*, is the second most popular distractor, with 5,3%[1]. We could see here an instance of the mirror-image principle, these children having mixed the graphemes *d* and *b*. We advance this with caution, since there are other things also that distinguishes the distractor

[1] Missing answers are included in the calculation of the popularity of each option.

from the key. The end-aversion principle could also explain in part the relative popularity of this distractor.

The second best performing item in tier 2, *hytte* (discrimination: 1,6) had a possible instance of another visual principle, the word-image principle:

bytte	hytte	hylle	hutre

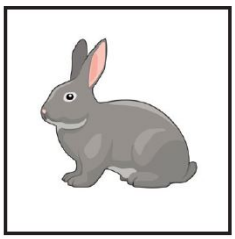
Option 3 in this item was developed with the word-image principle in mind. However, only 1,2% chose that option, weakening this principle for the 70%-90% ability range. Option 2 in the previous item, *drepe*, is then more convincing with 6,8% of the choices, although this could also be due to the phonetic similarity between the /e/ of *drepe* and the /o/ of *dråpe.*

The tendency to read only the beginning of the word could have trapped some children in the item *traktor* (discrimination 1,5):

trakter	kraftfor	traktor	faktor

8,1% chose this option, a very high popularity for a distractor. However there could be several reasons for this popularity. *Traktor* is a complex word, with two consonant clusters. An even more plausible explanation is the phonetic proximity between the /ə/ of *trakter* and the /u/ of *traktor*.

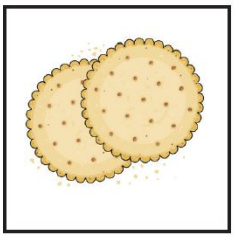
Visual principles of item development do not seem very effective in the tier 2 items. Moving on to the best performing items in tier 1, those that contribute to identifying the 10% weakest readers, they seem to have their place. All of the four following items from tier 2 can be explained by the word-image principle or the beginning-only principle:



kanon	kanne	kanin	kamin

(Kanin, discrimination: 2,9[2]*)*

[2] Items in tier 1 consistently discriminate better than items in tier 2, and the threshold for qualifying for the pile of best performing items is thus much higher in tier 1.



kjeft	kjeks	kjelke	kjele

(*Kjeks*, discrimination: 2,4)

øre	øm	orm	ørn

(*Ørn*, discrimination: 2,2)

kiv	vink	kvin	kniv

(*Kniv*, discrimination: 2,0)

1,5 % chose *kanon*, 2% chose *kjeft*, 1,6% chose *kjelke*, 2,5% chose *øre* and 3,8% chose *kiv.* Of course, other principles could contribute to explaining some of this, such as the complex grapheme *kj* and the consonant cluster *ks* in *kjeks*, or the consonant cluster *kn* in *kniv.* However, it seems, at least, that the beginning-only and word-image principles are strengthened for tier 1 items.

Principles concerning the phonetic manifestation of the sign.

Concerning phonetic principles, we have already examined two possible instances in tier 2 of the vowel change principle: *drepe* for *dråpe* and *trakter* for *traktor*. In tier 1, we also have an item where this seems the most plausible explanation, *lår* (discrimination: 2,0). Here, *lar* is the most popular distractor with 2,7% in position 3.

Some more specific phonetic problems seem characteristic for the two ability ranges. In tier 2 this is the confusion between /y/ and /ʉ/, as seen in the item already shown *hytte* and in *skute* (discrimination: 1,4):

skyte	snute	skute	skrue

In *hytte*, *hutre* is the most popular distractor with 8,6%, which is particularly impressive as it is in position 4. Of course, /y/ vs. /ʉ/, is not the only difference between the two, and there may also be some semantic leakage here (*hutre,* meaning *to shiver* may be attractive by association because of the snowy picture). Even so, it seems unquestionable that the u/y principle is at play here. This is even more obvious in *skute* where the distractor *skyte* attracted 17 % of the answers. The u/y principle thus seems to be a good indicator in tier 2.

In tier 1, we do not see that principle at play, but we find the neighbouring principle i/y, in the item *by* (discrimination: 2,0):

bi	bry	by	bu

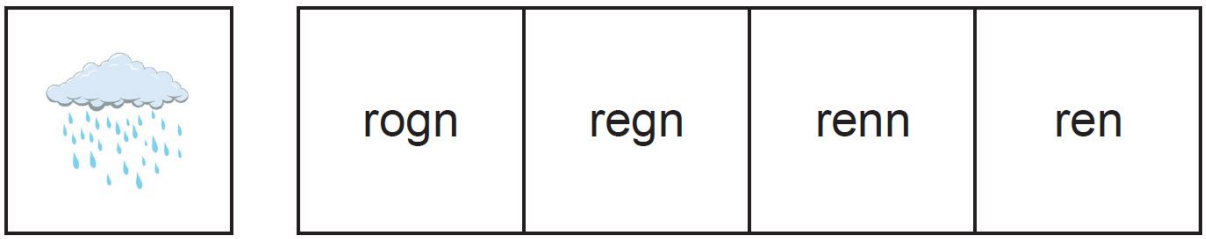
5 % of the children chose *bi*, strenghtening the i/y principle for this ability range. The number of occurrences here is obviously too small to be able to affirm the validity of the i/y principle and the u/y principle, let alone affirming that one is effective in tier 1 and the other one in tier 2, but these examples show at least that the principles can function.

Another important phonetic principle is the consonant cluster principle. We have already mentioned this as a complicating factor in *dråpe*, *traktor*, *skute*, *kjeks* and *kniv.* It is only in *dråpe* and *kniv* that we have included distractors that specifically tap the simplification of these clusters. In *dråpe*, *døpe* attracts 3,6% of answers, which is quite strong in position 3. However, this example may be corrupted by the vowel change. In *kniv*, *kiv* attracts 3,8%, by far the most popular distractor. This is a “clean” instance of simplifying a consonant cluster.

Mixed principles

Some principles are difficult to attribute only to either phonetic or visual difficulties, and two of these are at play among the best performing items. One is the permutation principle, positing that struggling readers will tend to choose distractors contain the same graphemes but where they have changed place. Only one of these items made it to the piles of best performing items: *kniv*. The two permuted distractors, *vink* and *kvin* attracted only 1,0% and 1,2% of responses respectively, which is quite low even for a tier 1 item. The soundness of this principle could seem doubtful.

The other mixed principle observed is the orthophonic spelling principle, positing that in words that do not have a completely orthophonic spelling in Norwegian, some readers will regularize the spelling. Only one markedly non-orthophonic word made it to the list of best performing items, *regn* (discrimination: 1,4):



This item ended up in tier 2, and was thus quite difficult. *Regn* is pronounced /ræin/ and there was no distractor that tapped the exact orthophonic spelling. However, options 3 and 4 come close, only simplifying the diphtongue. They attracted 5,9% and 5,4% of responses respectively. It is no surprise that an item based on the orthophonic spelling principle is difficult, and this principle may be considered unsound in a screening test, because over-orthophonic spelling can be seen as a normal step towards full literacy.

Conclusion

This preliminary analysis has corroborated some of the principles for item development in screening tests with which we started out. Both the general end-aversion principle and some of the subject-matter specific principles seem to function well. There also seems to be qualitative difference between tier 1 and tier 2 items.

However, this analysis is superficial in that it has considered only the items that performed best in the pilot. It needs to be refined by considering also other well-performing items, and also by comparing with items that did not perform well for the purpose – either discriminating badly or being too difficult (or too easy).

They also need to be tested more qualitatively by observing and listening to children solving the items while thinking aloud. Both of these further steps will be done before presenting the final paper in the conference.

Beyond the conference, it would be interesting to validate the principles further by a more controlled pilot, where only one principle is at play in each item, and where the key is evenly distributed among items.