JLTA 2011 Keynote Speech

Releasability of Language Test Specifications

Fred DAVIDSON

(University of Illinois at Urbana-Champaign, USA)
Releasability of Language Test Specifications

Fred DAVIDSON
University of Illinois at Urbana-Champaign, USA

Abstract
A test specification (‘spec’) is a generative document from which equivalent language test items or tasks can be produced. There are many formats for specs, but all share two elements: sample(s) of the items/tasks and guiding language that describes the sample(s). Through consensus-building and feedback, specs evolve and stabilize. A complete illustrative test spec is presented, based on workshops held in Japan in late 2011. A new problem in spec-driven test development is posed in this paper: releasability, which refers to whether a spec should be shared outside of the test development team, and if so, when and in what form. The illustrative spec is again used to explore releasability. A number of theoretical questions are posed about spec release, and future research about spec release is encouraged.

Keywords: specification, releasability

1. Language Test Specifications

A test specification (or “spec”) is a blueprint from which many equivalent test items or tasks can be produced. Specs are an important tool for the standardization and control in many testing systems. They are also a way to foster dialogue and consensus-building among educators. A spec contains guidance for writing test items or tasks, and such guidance helps ensure equivalence even when the precise content of items differs. Li (2006) presents a conceptual diagram for specs and their relationship to test validity, shown here in Figure 1.

1 This paper is based on the keynote presentation at the 15th Annual Meeting of the Japan Language Testing Association, Osaka, 29 October 2011.

2 Readers who wish to further discuss this article can reach Fred Davidson at fgd@illinois.edu
Figure 1 presents a part of a test spec that is (arguably) about half-way toward its destiny (see Appendix 1 for the whole test spec). It began as a demonstration spec in Davidson (forthcoming) and was revised in two workshops held in Japan in October of 2011 at Kansai University and at the East Shikoku Chapter of the Japan Association of Language Teachers, in Kochi. The marginal comments at the right are written by this author, who also facilitated both workshops. These marginal comments synthesize the critical feedback and consensus achieved at both workshops.
**Spec Title:** Testing Vocabulary in Context [("VIC")], ver. [v0.50] [with marginal comments]

**Guiding Language:**

This spec is intended for use in an ESL/EFL teaching situation in which the students need to figure out the meaning of words from contextual cues. A cue is a meaning-bearing element taken from elsewhere in the passage and/or from accompanying pictures and which is directly related to the word being tested. The word “context” thus refers to the passage and its accompanying pictures (if any) and to no other sources. The output of this spec is a multiple choice item, in which the selection of the correct answer depends on information found elsewhere in the body of text, the pictures, or both.

The stem of the item should follow one of these set formulas:

1. As used in line XXX, the word "YYYYY" means:
2. As used in line XXX, the word "YYYYY" refers to:
3. As used in line XXX, the word "YYYYY" is closest in meaning to:

The best choice should be directly discernable from the text, and should be given as choice (c) Choice order will be randomized before assembly of the full test.

---

**Figure 2.** A part of a Sample Specification. The whole version can be seen in Appendix 1.

Specs have two basic components: guidance about item writing and at least one sample item. In this spec, those two components are presented with very simple headers: “Guiding Language” and “Sample”. The Sample shows one test task that this spec is intended to produce. The rest of the spec is guidance about how to write the sample.

---

3 There are many organizational designs to write a spec. Davidson and Lynch (2002) present a method that is based on the work of Popham (1978) and which breaks down guiding language into sub-parts. However, all formats of test specs share two elements in common: sample(s) of the item/task that the spec will produce and guiding language that explains its rationale, resources, scoring rubric (if it is a rated item) and other guidance.
The example in Figure 2 is a detailed spec, which means that item writers are given quite a bit of advice and guidance about how to write the items. Typically, a large multi-item test is designed on a smaller number of detailed specs. For instance, a 100-item test might have ten or twelve detailed specs behind it, each generating some seven to ten items. The relationship of a detailed spec to the overall test is best summarized as shown in Figure 3, which presents a hypothetical “Table of Test Specs”. Each row in the table of specs would require one detailed spec; that is to say, this particular spec table would necessitate three detailed specs: one for vocabulary in context, one for skimming, and one for scanning. The vocabulary-in-context spec is presented as an example in this article.

<table>
<thead>
<tr>
<th>skill tested</th>
<th>Number of tasks/items per skill</th>
<th>Type of task/item per skill</th>
<th>Desired weighting</th>
<th>Special materials &amp; comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>vocabulary in context</td>
<td>20 items</td>
<td>multiple-choice with formulaic stem</td>
<td>50%</td>
<td>Passages needed; 2-3 items per passage</td>
</tr>
<tr>
<td>scanning</td>
<td>20 items</td>
<td>multiple-choice: specific information from passage</td>
<td>25%</td>
<td>passages needed; 2-3 items per passage</td>
</tr>
<tr>
<td>skimming</td>
<td>20 items</td>
<td>multiple-choice: general information from passage</td>
<td>25%</td>
<td>different passages from scanning. 1-2 items per passage</td>
</tr>
</tbody>
</table>

*Figure 3. A Table of Test Specifications (adapted from Davidson and Lynch, 2002).*

2. Spec Theory

A test specification is best seen as an evolutionary document, and Figure 2 attempts to capture a particular spec at a particular point in its evolution. The classic generative nature of specs does not arise immediately. It depends on feedback (see Figure 1). Typically, this feedback comes from critical discussion amongst the test development team, as shown in
the marginal comments in Figure 2. It can also come from tryouts of the test items or tasks, if resources permit.

Another major element of spec theory concerns their origins. How are specs created in the first place? There seem to be two possibilities. First, a spec might be created by analysis of an existing test task; for example, the VIC item in Figure 2 may exist in the files of a teaching institution, and that institution may want to create equivalent tasks for a new version of its test. However, the institution has no written detailed spec. This is ‘reverse engineering’ (Davidson and Lynch, 2002: 41; Fulcher and Davidson, 2007: 56-58), which is when the test development team uses an existing item or task as the sample and then develops guiding language around it. Reverse engineering often has one of two functions. First is straight reverse engineering, which is the creation of equivalent tasks without changing them or improving them. Second is critical reverse engineering, which often begins with the straight variety but usually has the larger goal to improve the test in fundamental ways. If the sample VIC item in Figure 2 came from an actual test, then the marginal comments reveal that sample is already subject to change (e.g. the critical remark about the word ‘pedal’).

The second origin of test specs is an archetype (Fulcher and Davidson, 2007: 58-59), which is a typical way to test a particular skill. Figure 4 presents some guiding language: What kind of an item fits the guiding language in Figure 4? The reader is encouraged to consider that question. The actual test item from which it was (straight) reverse engineered appears in Appendix 2 to this paper. Very likely, the item in the appendix matches or closely resembles the one the reader expects to see. Language test development may fall into typical and familiar archetypes.

The purpose of this item type is to test reading of a descriptive statistical table. The statistical report is from a scientific survey or study, and it refers to common everyday knowledge. There is a what-question about the relationship of two or more variables in the data table. There are five multiple-choice responses. One choice is correct. The other four choices – the distracters – are wrong, because they mis-interpret the report very slightly. Close reading is required to distinguish between the choices.

Figure 4. Guiding language to match the sample item shown in Appendix 2.

In summary, the elements of spec theory are as follows:
1. Specs fulfill a classic function: they control equivalency of items and tasks.
2. Specs exist.
   2a. Reverse engineering is fundamental.
   2b. Archetypes exist.
3. Specs evolve, and thus they contribute evidence of test validity.
4. The specs (and the test they generate) are not launched until ready.
5. Discussion happens and that leads to transparency.
6. And this is discussion to which all are welcome.
   (adapted from Fulcher and Davidson, 2007, pp. 60-61)

The final three elements of the theory are developmental ideals or, possibly, indicators of quality control. For example, the illustrative spec in Figure 2 is clearly not ready to generate tasks yet. Its marginal comments reveal a serious concern about internal elements of its design and about its sample item. The teams that worked on this spec (at the workshops in Japan) felt that it has promise, and that further work on this spec is merited. The reader is left to his or her own personal judgment about its basic quality, but at a minimum, the reader may enjoy presenting this version of that spec to colleagues for discussion and further evolution. A clean copy is presented in Appendix 3.

3. “We don’t have any specs”

One element of spec theory deserves special mention: point (2) asserts that specs exist. This author has often heard a familiar complaint: “We don’t have specs.” However, when teachers (who make that complaint) work together and analyze their tests, they see that there are common styles and similar approaches to testing. For instance, if several teachers have items that test reading of statistical tables, they might have something similar to the task shown in Appendix 2. That is an example of an archetype. Furthermore, the teachers can discuss and articulate what they do for that kind of task. They can examine their reading items and can explain how their items are written. That is the genesis of guiding language. When the existing samples are combined with this early guiding language, we see that a spec in fact exists. Perhaps the spec has existed all along, but it was (simply) never written down; it was a ‘spec-in-the-head’ (Davidson and Lynch, 2002, p. 65).

Why write it down?

There are two reasons. First, writing down a spec helps to train newcomers to the teaching situation, because new teachers can see how things are typically done. Second, as all teachers (old and new) look at the spec, they can discuss its strengths and weaknesses. What began as straight reverse engineering can become critical reverse engineering, and teachers can work together to alter their testing as they see fit. Discussion about item quality is more productive if it is about a group of similar items and not about one single question, and by grouping items together according to their spec, test improvement has a much broader scope.
4. Releasability

As a spec evolves, it reflects a massive investment by the test development team. The team first creates the spec, perhaps by reverse engineering, and very likely by archetypes. The spec is debated and changed. It evolves through different versions. Feedback affects each version in creative ways, and at some point, the team decides to make a critical investment, if resources permit: to try out the spec’s sample task with a group of representative test takers. The spec evolves further, and if resources still permit, perhaps it is piloted again. Eventually – if the spec survives – it becomes a productive and generative document. It may only be a few pages long, but its length is the tip of an iceberg of investment.

What if this is a high-stakes test, such as an entrance and placement exam developed by a particular college or university? Test-takers (and their families) want to know what the test measures. They might have access to some former sample items, but it would be more helpful if they had both samples and some guiding language. Perhaps they ask the university to give them such full and rich information.

Alternatively, the test development group may be a commercial publishing company. The company has invested a lot of time and money to build a bank of new test specs, and from that bank, it is generating new items for a new testing product. Competitors are interested in this new product and want to see its generative blueprint.

Human nature may kick in and the team may become unwilling to share the spec with other testing situations, if for no reason than these: (to competitors) “Why should those other groups benefit from our long hard work?” (and to test-takers) “Why should we give them our full spec? Doing so would help them to cram for the test without fuller learning.”

The problem at hand is “releasability”, which is the degree to which test specifications are shared outside of the test development team. If a test team has high releasability, then they share many or perhaps all of their specs. If releasability is low, that means the generative blueprint for the test is secret and guarded.

There seem to be two logical problems with release of test specs: timing and the nature of the released material.

5. Timing of release

Logically, if a testing team releases its spec, there is the question: when should that be done? And, should the team release an exact version of its specs, or should it edit the specs to remove some information? Should the team have two parallel versions of its specs – an internal-use version and a releasable version – and if so, at what point in test development
should this happen? It is assumed that the testing team would not wish to release a spec until it is relatively stable; for example, the VIC spec discussed earlier in this article (about half-way toward its destiny) is probably not yet releasable in any form.

6. What might a releasable spec look like?

In order to discuss releasability fully, a more evolved spec is needed – and so the VIC spec in Figure 2 must evolve a bit more. That spec still has some serious problems, which workshop participants noted. These problems were placed into a “waiting room”, which is a separate document (or element of a spec) showing details and problems that will need to be solved, as follows:

(1) To avoid mismatch between the sample items and guiding language, we should refine further the guidance about distracters.

(2) We need a much better operational definition of “plausible”. Once obtained, this definition should appear at the opening of the spec, alongside the discussion of “cue” and the operational definition of “context”.

(3) Other words needing further definition and debate are: “intermediate” and although some revision was done on it here, “context”.

(4) Is the handcycle passage appropriate? How much specialized knowledge is needed to ensure that test-takers will, in fact, seek meaning elsewhere in context and not rely on real-world knowledge? A good way to approach this problem would be to work up some other possible passages – to define ‘specialized’ only after a range of example texts are available for comparison and discussion.

(5) We will probably have to figure out the difference (in the stem formula) between “mean”, “refers to”, and “is closest in meaning to”.

(6) We haven’t discussed the pictures yet, at least not much.

Workshop participants noted in comment (5) that these words are confusing: “mean”, “refers to” and “is closest in meaning to”.

Perhaps by version 0.80, the developers prefer “is closest in meaning to”. Furthermore, they have done some analysis of online word lists and thesaurus websites (like http://thesaurus.com/). By version 0.80 of the spec, perhaps the team has expanded its guiding language to read as follows: “Closeness of meaning should be judged by study of an online thesaurus like http://thesaurus.com/, and for any tested word (such as “assembly”) each item writer shall include a screen image of the intended meaning.” By this point, the spec includes a sample screen image from a thesaurus website to illustrate the precise and intended meaning of the target test word, using the search feature of the online thesaurus. By searching for “assembly” and checking other words associated with it, the test developer eventually finds “construction” (the intended meaning) and sees that its
entry includes “assembly”. The thesaurus verifies the intended meaning. This is shown in Figure 5, which is the entry for the intended meaning, showing the target word in bold.

Main Entry: **construction** [kuhn-struhk-shuhn] **IPA**

Part of Speech: *noun*

Definition: creation, building

Synonyms: architecture, arrangement, **assembly**, build, cast, composition, conception, constitution, contour, cut, development, disposition, edifice, elevation, erecting, erection, fabric, fabricating, fabrication, figuration, figure, form, format, formation, foundation, improvisation, invention, making, manufacturing, mold, origination, outline, plan, planning, prefab, prefabrication, putting up, raising, rearing, roadwork, shape, structure, system, systematization, turn, type

Antonyms: destruction, disfigurement, disorganization, ruin, ruins

*Figure 5*. A thesaurus entry for “construction” which is the intended meaning for “assembly” as used in the test passage. Downloaded from [http://thesaurus.com](http://thesaurus.com).

It is probably best not to tell the public that the test developers consult a particular website. Doing so might affect the study habits for the test, because test-takers would (simply) try to memorize the various entries at that website. The released version would state simply that target words are assessed by multiple choices that are “closest in meaning to” the target; however, the non-released (or “internal”) spec would include the guidance to item writers shown here: it would state a particular website to consult and would ask the item writer to show the intended meaning.

In short, production of a releasable version of a test spec involves careful editing of the internal spec to remove details that affect study for the test, test security, and other considerations that might alter the test’s validity.

7. **Closing remarks: some questions of theory about releasability**

Several areas of research suggest themselves:

1. **Scope and focus**: how much of a spec should be released, and why?
These questions are provocative and far-reaching. All of these issues can be addressed productively if test developers and researchers start small; perhaps they could attack one single element of this theory – ‘focus’. For instance, is it (really) a bad idea to tell the wider public what websites are consulted during language test development? Would it really be harmful to let students know that test developers regularly use a website like the thesaurus link, shown above? In fact, maybe releasing that point actually triggers positive washback on instruction and teaching. This paper is offered as a way to stimulate discussion and collaboration amongst language testers in Japan so that the above questions can be explored.

References


Appendix 1. A Sample Specification. Adapted from Davidson, F. (forthcoming) "Test Specifications." The Encyclopedia of Applied Linguistics (published by Blackwell), Carol Chapelle, General Editor. See below for all the comments in Appendix 1.

Spec Title: Testing Vocabulary in Context ("VIC"), ver. 0.60 [with marginal comments]

Guiding Language:

This spec is intended for use in an ESL/EFL teaching situation in which the students need to figure out the meaning of words from contextual cues. A cue is a meaning-bearing element taken from elsewhere in the passage and/or from accompanying pictures and which is directly related to the word being tested. The word "context" thus refers to the passage and its accompanying pictures (if any) and to no other sources. The output of this spec is a multiple choice item, in which the selection of the correct answer depends on information found elsewhere in the body of text, the pictures, or both.

1. The stem of the item should follow one of these set formulas:
   1. As used in line XXX, the word "YYYYY" means:
   2. As used in line XXX, the word "YYYYY" refers to:
   3. As used in line XXX, the word "YYYYY" is closest in meaning to:

The best choice should be directly discernable from the text, and should be given as choice (c) Choice order will be randomized before assembly of the full test.

Comment [fgd1]: Test development projects often use their own abbreviations, such as "VIC".

Comment [fgd2]: In spec evolution metaphor (Figure One), the operational deliverable spec is version 1.0. I feel this spec is coming along. We are a bit more than halfway there. My experience is that as we move along, we encounter and fix problems of greater difficulty. 0.60 feels about right as the current version number, metaphorically.

So that all my comments can be seen, I have split the guiding language across two pages. I then present a clean version of 0.60 without marginal comments: ver 0.61. In spec based test development, you create such numbering standards: record commentary in one version, raise the version number slightly, and then save it in a "clean" copy.

Either way, the numbering allows you to (later) "audit" the development of the spec as Li did in her 2006 M.A. thesis. She compared different versions of a set of specs to see how difficult problems were resolved and refined.

Comment [fgd3]: This is my attempt to operationally define context, as real-world knowledge part of "context"? Here, I take the stance that it is not — that the definition of context shall be limited to the text and to its supportive pictures. I should also note the phrase: "...and to no other sources", which is an example of negative guiding language. Often, in spec development, we need to say both what we want and what we do not want, both are helpful.

Comment [fgd4]: A multiple-choice item has four parts: the stem (the governing question or statement), "the choices" (here four), and in the choices, one is best/correct, "the key" and the others are not, "the distractors".

This change allows greater flexibility and creativity by the item writer. The contextual cue is expanded — it need not be only direct meaning, but it could be referential, or close meaning.

My instinct is that this is a good change, but that down the road in later versions, we will face a challenge about the difference in the three words: "mean", "refer to" and "closest in meaning".

Perhaps it is best to settle on one or the other, or more likely, it is better to define each precisely so that the creative scope of the spec is better understood. I've put this in the Waiting Room.
Distracters (a) and (b) should be a known and plausible meaning of the target word but one that is irrelevant to the text as presented. Distracter (d) should refer to word or concepts elsewhere in the text but present an implausible meaning of the target word.

Typically, two to three items per text will be generated. If a text can support more items, then that is permitted.

The text should be [100-200 words at a difficulty level appropriate for third-year college study of EFL]. The topic should be a familiar concept, thing, place, or person; however, some specialized knowledge or exposition is permitted, so that contextual vocabulary cues can be assured.

Supporting image(s) are permitted.


Comment [fged5]: We briefly discussed this word at Kochi. As I note in the footnote of this spec, it first appears in a forthcoming chapter that I have written. In that chapter, I take up the complex problem of defining “plausibility” with a series of examples, so that readers can form their own conclusions about its meaning.

Comment [fged6]: This point came up at Kochi. I think it is a good idea to broaden the guiding language here, because as we discussed at Kochi, there is a lot of text time used to read a passage – in fact, that text is what “spawned” the idea for a new spec, as discussed below.

Comment [fged7]: This may not be the best direction to pursue, but it is an attempt to better define “intermediate”. First, the length of the passage is more precisely limited. 50-200 is a broad range. Second, rather than a vague word like “intermediate”, the spec now cites a certain amount of prior study: at least two previous years of college EFL.

Comment [fged8]: Both the Kochi and Kansai groups debated whether handcycles (the topic of the sample) is (a) familiar enough, yet (b) specialized so as to allow new words to be tested. At Kochi, we decided to delay this decision – hence it is in the Waiting Room below.

Here’s what I think could happen: if this spec were to go through a few more versions, we could keep the handcycle passage, but we could further refine and develop difficult concepts like “plausible” and “context” and “intermediate”. Then we could return to the handcycle passage and see if it still works. If it does not, then later, we find a new passage. At that point, the spec may have to contain greater detail and guidance about how to craft the passage.

Personally, I think that the handcycle passage can survive, because testing vocabulary in context means we should have a passage topic that is not terribly familiar. If a passage topic is very familiar, then how can we ensure that test takers seek contextual cues from the passage and/or its pictures? The whole problem may become a matter of how the passage is written rather than its actual topic.

I did make one minor change in this paragraph: “obtained” became “assumed.”

Comment [fged9]: Neither workshop really discussed the pictures. I am certain that they will come up for later discussion, if this spec evolves.
Sample:

1 A handcycle is a type of human powered land vehicle powered by
2 the arms rather than the legs, as on a bicycle. Most handcycles
3 are tricycle in form, with two coasting rear wheels and one
4 steerable powered front wheel.

5 Many manufacturers have designed and released hand-powered
6 recumbent trikes, or handcycles. Handcycles are a regular
7 sight at HPV meets and are beginning to be seen on the streets.
8 These usually follow a delta design with front wheels driven by
9 standard derailleur gearing powered by hand cranks. Brakes levers
10 are usually mounted on the hand holds, which are usually set
11 with no offset rather than the 180° of pedal cranks. The entire
12 crank assembly and the front wheel turn together, allowing the
13 rider to steer and crank simultaneously.

14 Thanks to modern technology, handcycles come in a variety of
15 styles, making them accessible to people of all abilities,
16 including many persons with disabilities. There are also
17 hybrids between a handcycle, a recumbent bike and a tricycle.

Comment [fgd10]: Please see the change to the second sample item below. I am not certain if we should remove "pedal" here, as well. Perhaps. Perhaps not. It may not matter.
As used in line 7, the word "HPV" means:
(a) hand project version
(b) hand propelled variation
(c) human-powered vehicle
(d) hybrid-propelled vehicle

As used in line 12, the word "assembly" means:
(a) a gathering of people
(b) a gathering of bicycles
(c) a handcycle's crank mechanism
(d) a bicycle's crank mechanism

Comment [fgd11]: This attempts to make (d) less attractive. However, I am not sure that it fits the distractor mode in the guiding language above. See the Waiting Room for more thoughts about the distractors, which takes on the matter of match between the sample and the guiding language.

Comment [fgd12]: As suggested by the Kochi workshop. They suggested this change because the word "pedal" derives from "ped" which means foot. The point is: handcycles call the things you hold (with your hand) a "pedal" even though the feet are not involved. We can easily avoid this bit of odd bicycle jargon by using "crank" here.
## Comments

<table>
<thead>
<tr>
<th>fgd1</th>
<th>Test development projects often use their own abbreviations, such as ‘VIC’.</th>
</tr>
</thead>
<tbody>
<tr>
<td>fgd2</td>
<td>In spec evolution metaphor (Figure One), the operational/deliverable spec is Version 1.0. I feel this spec is coming along. We are a bit more than halfway there. My experience is that as we move along, we encounter and fix problems of greater difficulty. 0.60 feels about right as the current version number, metaphorically. So that all my comments can be seen, I have split the guiding language across two pages. I then present a clean version of 0.60 without marginal comments: ver 0.61. In spec-based test development, you create such numbering standards: record commentary in one version, raise the version number slightly, and then save it in a “clean” copy. Either way, the numbering allows you to (later) “audit” the development of the spec, as Li did in her 2006 MA thesis: she compared different versions of a set of specs to see how difficult problems were resolved and refined.</td>
</tr>
<tr>
<td>fgd3</td>
<td>This is my attempt to operationally define context. Is real-world knowledge part of “context”? Here, I take the stance that it is not – that the definition of context shall be limited to the text and to its supportive pictures. I should also note the phrase: “…and to no other sources”, which is an example of negative guiding language. Often, in spec development, we need to say both what we want and what we do not want. Both are helpful.</td>
</tr>
<tr>
<td>fgd4</td>
<td>A multiple-choice item has four parts: “the stem” (the governing question or statement), “the choices” (here four), and in the choices, one is best/correct: “the key” and the others are not: “the distracters”. This change allows greater flexibility and creativity by the item writer. The contextual cue is expanded – it need not be only direct meaning, but it could be referential or close meaning. My instinct is that this is a good change, but that down the road in later versions, we will face a challenge about the difference in the three words: “mean”, “refer to” and “closest in meaning”. Perhaps it is best to settle on one or the other, or more likely, it is better to define each precisely so that the creative scope of the spec is better understood. I’ve put this in the Waiting Room.</td>
</tr>
<tr>
<td>fgd5</td>
<td>We briefly discussed this word at Kochi. As I note in the footnote of this spec, it first appears in a forthcoming chapter that I have written. In that chapter, I take up the complex problem of defining “plausibility”. I don’t try to resolve it here, but I have placed this issue in the Waiting Room, below.</td>
</tr>
</tbody>
</table>

---

- 17 -
fgd6 This point came up at Kansai. I think it is a good idea to loosen the guiding language here, because as we discussed at Kochi, there is a lot of test time used to read a passage – in fact, that fact is what “spawned” the idea for a new spec, as discussed below.

fgd7 This may not be the best direction to pursue, but it is an attempt to better defining ‘intermediate’. First, the length of the passage is more precisely limited. 50-200 is a broad range. Second, rather than a vague word like “intermediate”, the spec now cites a certain amount of prior study: at least two previous years of college EFL.

fgd8 Both the Kansai and Kochi groups debated whether handcycles (the topic of the sample) is (a) familiar enough, yet (b) specialized so as to allow new words to be tested. At Kochi, we decided to delay this decision – hence it is in the Waiting Room below.

Here’s what I think could happen, if this spec were to go through a few more versions: We could keep the handcycle passage, but we could further refine and develop difficult concepts like “plausible” and “context” and “intermediate”. Then we could return to the handcycle passage and see if it still works. If it does not, then later, we find a new passage. At that point, the spec may have to contain greater detail and guidance about how to craft the passage.

Personally, I think the handcycle passage can survive, because testing vocabulary in context means we should have a passage topic that is not terribly familiar. If a passage topic is very familiar, then how can we ensure that test-takers seek contextual cues from the passage and/or its pictures? The whole problem may become a matter of how the passage is written rather than its actual topic.

I did make one minor change in this paragraph: “obtained” became “assured”.

fgd9 Neither workshop really discussed the pictures. I am certain that they will come up for later discussion, if this spec evolves.

fgd10 Please see the change to the second sample item below. I am not certain if we should remove “pedal” here, as well. Perhaps. Perhaps not. It may not matter.

fgd11 This attempts to make (d) less attractive. However, I am not sure that it fits the distracter model in the guiding language above. See the Waiting Room for more thoughts about the distracters, which takes on the matter of match between the sample and the guiding language.

fgd12 As suggested by the Kochi workshop. They suggested this change because the word “pedal” derives from “ped” which means foot. The point is: handcycles call the things you hold (with your hand) a “pedal” even though the feet are not involved. We can easily avoid this bit of odd bicycle jargon by using “crank” here.
Appendix 2. This is the sample item for Figure 4. Note. The temperatures are given in Farenheit.

Sample Item:

Which represents the relationship between temperature and latitude?

<table>
<thead>
<tr>
<th>City</th>
<th>Temperature</th>
<th>Latitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bismark</td>
<td>64.3°F</td>
<td>46.8°</td>
</tr>
<tr>
<td>Sioux Falls</td>
<td>68.4°F</td>
<td>43.5°</td>
</tr>
<tr>
<td>Omaha</td>
<td>73.0°F</td>
<td>41.2°</td>
</tr>
<tr>
<td>Kansas City</td>
<td>76.1°F</td>
<td>39.1°</td>
</tr>
<tr>
<td>Wichita</td>
<td>73.6°F</td>
<td>37.7°</td>
</tr>
<tr>
<td>Dallas</td>
<td>82.0°F</td>
<td>33.2°</td>
</tr>
</tbody>
</table>

A. As latitude increases, temperature decreases.
B. As latitude increases, temperature increases.
C. As latitude decreases, temperature decreases.
D. As latitude decreases, temperature remains constant.
E. As latitude increases, temperature remains constant.

Source: IMAGE Mathematics Sample Items Grade 11
Appendix 3. This is a clean copy of the spec shown in Figure 2 – the reader is encouraged to discuss this spec with colleagues and evolve it further.

Spec Title: Testing Vocabulary in Context ("VIC"), ver. 0.61 [same as 0.60 but without the marginal comments]

Guiding Language:

This spec is intended for use in an ESL/EFL teaching situation in which the students need to figure out the meaning of words from contextual cues. A cue is a meaning-bearing element taken from elsewhere in the passage and/or from accompanying pictures and which is directly related to the word being tested. The word "context" thus refers to the passage and its accompanying pictures (if any) and to no other sources. The output of this spec is a multiple choice item, in which the selection of the correct answer depends on information found elsewhere in the body of text, the pictures, or both.

The stem of the item should follow one of these set formulas:

1. As used in line XXX, the word “YYYYYY” means:"
2. As used in line XXX, the word “YYYYYY” refers to:"
3. As used in line XXX, the word “YYYYYY” is closest in meaning to:"

The best choice should be directly discernable from the text, and should be given as choice (c) Choice order will be randomized before assembly of the full test.

Distracters (a) and (b) should be a known and plausible meaning of the target word but one that is irrelevant to the text as presented. Distracter (d) should refer to word or concepts elsewhere in the text but present an implausible meaning of the target word.

Typically, two to three items per text will be generated. If a text can support more items, then that is permitted.

The text should be 100-200 words at a difficulty level appropriate for third-year college study of EFL. The topic should be a familiar concept, thing, place, or person; however, some specialized knowledge or exposition is permitted, so that contextual vocabulary cues can be assured.

Supporting image(s) are permitted.

Sample:

1 A handcycle is a type of human powered land vehicle powered by the arms rather than the legs, as on a bicycle. Most handcycles are tricycle in form, with two coasting rear wheels and one steerable powered front wheel.

5 Many manufacturers have designed and released hand-powered recumbent trikes, or handcycles. Handcycles are a regular sight at HPV meets and are beginning to be seen on the streets.

8 These usually follow a delta design with front wheels driven by standard derailleur gearing powered by hand cranks. Brake levers are usually mounted on the hand holds, which are usually set with no offset rather than the 180° of pedal cranks. The entire
12  crank assembly and the front wheel turn together, allowing the
13  rider to steer and crank simultaneously.

14  Thanks to modern technology, handcycles come in a variety of
15  styles, making them accessible to people of all abilities,
16  including many persons with disabilities. There are also
17  hybrids between a handcycle, a recumbent bike and a tricycle.

As used in line 7, the word “HPV” means:
(a) hand project version
(b) hand propelled variation
(c) human-powered vehicle
(d) hybrid-propelled vehicle

As used in line 12, the word “assembly” means:
(a) a gathering of people
(b) a gathering of bicycles
(c) a handcycle’s crank mechanism
(d) a bicycle’s crank mechanism
Waiting Room:

(1) To avoid mismatch between the sample items and guiding language, we should refine further the guidance about distracters.

(2) We need a much better operational definition of “plausible”. Once obtained, this definition should appear at the opening of the spec, alongside the discussion of “cue” and the operational definition of “context”.

(3) Other words needing further definition and debate are: “intermediate” and although some revision was done on it here, “context”.

(4) Is the handcycle passage appropriate? How much specialized knowledge is needed to ensure that test-takers will, in fact, seek meaning elsewhere in context and not rely on real-world knowledge? A good way to approach this problem would be to work up some other possible passages – to define ‘specialized’ only after a range of example texts are available for comparison and discussion.

(5) We will probably have to figure out the difference (in the stem formula) between “mean”, “refers to”, and “is closest in meaning to”.

(6) We haven’t discussed the pictures yet, at least not much.