An Application of Reverse Engineering to Automatic Item Generation:

A Proof of Concept Using Automatically Generated Figures

William A. Lorié

Questar Assessment, Inc.

April 2013

Abstract

A reverse engineering approach to automatic item generation (AIG) was applied to a figure-based publicly released test item from the Organisation for Economic Cooperation and Development (OECD) Programme for International Student Assessment (PISA) mathematical literacy cognitive instrument as part of a proof of concept. The author created an item template from which three items were randomly generated from within each of six types defined by a feature deemed to be most likely to affect item difficulty, for a total of eighteen distinct items. To assess their equivalence, these items were embedded in otherwise identical test forms and administered to human intelligence task workers on the Amazon Mechanical Turk system. One level of the type-defining feature appeared to affect item difficulty systematically. The author provides a task requirement rationale for removing this level. Implications for AIG theory and practice are discussed.

An Application of Reverse Engineering to Automatic Item Generation: A Proof of Concept Using Automatically Generated Figures

Testing programs often require the development of many new test items that mirror but are not identical to items in a reference test form. Test developers usually develop the new items by authoring new ones consistent with item writing guidelines and item and test specifications for the overall testing program. Although well supported, this process does not guarantee that the combination of new items will be equivalent in content or difficulty to the reference test form, an important goal of continued item development.

In this study, a different approach to new item development—*reverse engineering*—is applied in the context of automatic item generation (AIG; Gierl and Haladyna, 2013; Haynie, Haertel, Lash, Quellmalz, and DeBarger, 2006; Masters, 2010). The author presents a method for selecting an item exemplar, developing an item template for it, studying the properties of the items generated from that template, and revising the template for future deployment. As part of this proof of concept study, the exemplar item is chosen deliberately such that the item generation routine requires the automatic production of new figures with underlying properties similar to those in the exemplar.

Drawing upon concepts from engineering disciplines and writing from a test development perspective, Haynie et al. (2006) define reverse engineering as

> [T]he process of creating a design or blueprint by analyzing a final product or system—often via identification of system components and their interrelationships—and creating representations of that product or system in an enhanced form or at a higher level of abstraction. (p. 6)

When applied in the context of AIG, reverse engineering (Gierl & Haladyna, 2013; Masters, 2010) refers to the process of deriving an item-generating template from an exemplar item,

considered to have a high level of fidelity to the construct being assessed. In the case of complex constructs such as those that form the bases of large-scale assessments of student achievement, the exemplar item being modeled need only show fidelity to one of the many types of items referenced in test blueprints and specifications.

In the proof of concept study reported here, reverse engineering is applied both to create an item-generating template and to build upon the exemplar to better support the item's contribution to score-based inferences of respondents' knowledge and skills. Here the author proposes three essential elements for reverse engineering as applied to automatic item generation: (a) a justified exemplar, (b) a justified item schema, and (c) an item generator.

The first element of a reverse engineering is an exemplar item. Whether this item is identified or created, it should have a high level of fidelity to one aspect of the target domain (Kane, 2006). The exemplar's status as an exemplar for the target construct, and its candidacy for AIG modeling, should be justified. Task requirements or cognitive models to the extent possible and practical can be part of this justification, and can be based on principles of evidence-centered design (Mislevy, Steinberg, and Almond, 2002; see also Huff, Steinberg, and Matts, 2010; Huff, Alves, Pellegrino, and Kaliski, 2013) or assessment engineering (Lai and Gierl, 2013; Luecht, 2013; Luecht, 2007).

What constitutes a good judgment about an item's candidacy for AIG modeling is less well defined, but presumably rests on subject matter experts' being able to conceive of a suitably large number of alternative items that preserve some elements of the original item, and introduce variations that do not alter what the item measures. Importantly, it is not necessary that the alternatives generated by an AIG template be *independen*t: AIG does not assume that two or

more items generated from the same template must be eligible to be on the same test form or (in the case of adaptive forms of testing) be administered during the same test event.

An exemplar and its justification lead naturally to the development of a *schema*—here meant as a blueprint for building a specific generator. The important components of a justified item schema are specification of its fixed elements, its variable elements, and any dependencies between those elements. Since a schema provides the range of values that variables can have, it must supply a justification for those ranges. The term *item template*, as employed by Lai & Gierl (2013) in the context of assessment engineering, is synonymous with *schema* as used here.

The end goal of reverse engineering is to implement the item schema in a practical way—that is, to develop an item generator, also called a *template* in this study. The template implements the schema; items are the output.

The preceding describes reverse engineering elements at the level of specific *items*, not at the level of admissible test forms or item pools. The greatest benefit of reverse engineering, however, is at the aggregate level, where entire testing programs are modeled. For that goal, a reverse engineering approach requires two other elements: A form- or test-event- generation protocol for drawing upon the individual item generators, and a validity argument that supports the claim that scores earned on test events generated from that protocol are equivalent to those from the original set of (hand-written) forms or test events.

In this study, reverse engineering is applied to create, study, and refine an item template based on a publicly released test item from the Organisation for Economic Cooperation and Development (OECD) Programme for International Student Assessment (PISA) mathematical literacy cognitive instrument (OECD, 2006). The item, labeled M161Q01, was the only item in a unit named “Triangles” and was introduced in the 2000 administration of PISA.

An item template was derived from the target item through reverse engineering to generate multiple distinct items. Most of the features chosen for these variations were intended to introduce *incidental* changes—that is, alterations that should not induce a change in item difficulty, resulting in what are termed item *clones* or *isomorphs* (Bejar, 2002). One feature was deemed to be a potential *radical* (a feature that affects item difficulty; Irvine, 2002) due to variations in technical vocabulary among the different values of that feature. The feature was retained in the template to expand the template's range as well as to study the extent to which it might induce a significant variation in item difficulty.

Three items were randomly generated from within each of six *types* defined by this potentially radical feature, resulting in eighteen unique items, nested within the six types. Each of the items were embedded in a fixed position of a five-item test form assessing competence in geometry, and the eighteen distinct instruments were administered to human intelligence task (HIT) workers on the Amazon Mechanical Turk system (Amazon, 2012) to assess the degree to which the generated items can be considered equivalent.

The original purpose of this study was twofold: (a) to examine the feasibility of developing an item template for a figural item of an important testing program, using a reverse engineering approach, and (b) to assess the extent to which items generated from the template could be treated as randomly equivalent. The author deliberately chose a context in which figural elements required automatic generation to investigate the extent to which the approach could assist in eliminating the need for laborious art production for new items, as well as enhancing the construct definition of tests including such items.

Development of the automatic item-generating template for this study demonstrated the potential for generating increasingly complex (and less trivial) variations. Exploring non-trivial

variations enhances the contributions that a reverse engineering approach can make to AIG. Thus, the second component of the original purpose of this study was modified to address those specific non-trivial variations.

Method

Instrument Development

In this study, an item-generating template was developed and the performance of different instances of that item was examined. To compare item performance, the item instances were each incorporated as the second item of an otherwise fixed five-item test form, and the different resulting test forms were administered to randomly equivalent groups of human intelligence task (HIT) workers on a HIT system, who accepted the task of responding to the items for a fee consistent with other HIT tasks on the system. The design of this study is thus a balanced administration of eighteen distinct instances of the study item, with three instances for each of six item types. A sample HIT is displayed in Appendix A.

The study instrument consisted of eighteen test forms of five selected response items each. The items all assessed a variety of concepts typically taught in geometry at the secondary school level. To create a common context across conditions, the forms were made identical except for the second item, which in each form was a different item generated from a common template. Those eighteen items can be further subdivided into six groups of three items each, each group having a further feature in common: the way in which the first two constraints in the stimulus text were presented, as will be described in more detail in the section on the Triangles template.

Items 1, 3, 4, and 5 were written by the study author, and assessed understanding of lengths associated with a circle, angles created by intersecting two parallel lines with a third line,

the effects of enlarging figures, and the features of simple solids, respectively. Those four items were written specifically to sample a variety of related tasks assessing skills in geometry, and create a common, related context for the study item. All versions of Item 2 assessed a respondent's ability to match a written representation of a geometrical situation to a figural representation. All versions of Item 2 were generated from a template created by the study author, and modeled on PISA Item M161Q01, administered in 2000 and publicly released in 2006 (OECD, 2006).

The Triangles Template

PISA Item M161Q01 describes a right triangle, states where the right angle is located, states an inequality between the legs of the triangle, introduces two separate midpoints of two of the triangle's sides, introduces a sixth point and its relationship to the borders of the triangle, and states an inequality between the two line segments created by that sixth point and the two midpoints. Examinees are asked to indicate which of five figures, presented as answer options, fits the text description.

Construction of the Triangles template, based on PISA Item M161Q01, began with a task requirements analysis of M161Q01. The aim of this analysis was to uncover the task structure of M161Q01 and explore potential variations of the item that do not alter that underlying task structure. Building a representation of that task structure was a precursor to developing a schema for the template.

M161Q01 readily implies a *constraint satisfaction task*, in which several text-based constraints are presented, each of which must be met by (in this case) a figural depiction. Each of the figures presented as options in M161Q01, except for the keyed response, fails to meet at least

one constraint in the stimulus description, and the pattern of satisfaction can be captured in a constraint satisfaction matrix, presented in Table 1.

Table 1

Constraint Satisfaction Matrix for PISA Item M161Q01

Constraint	Figure A	Figure B	Figure C	Figure D	Figure E
1 – PQR is a right triangle	1	1	1	1	0
2 – The right angle of PQR is at R	1	1	0	1	0
3 – RQ is less than PR	0	1	1	1	0
4 – M is the midpoint of PQ	1	0	1	1	0
5 – N is the midpoint of QR	0	0	1	1	0
6 – S is inside triangle PQR	0	0	1	1	1
7 – MN is greater than MS	1	1	1	1	1

Note: "1" denotes satisfaction of the constraint; "0" denotes failure to satisfy. Figure D is the key.

It is useful to note that although the analysis of M161Q01 in this way may yield valuable information about the way in which respondents approach the item, the task requirements analysis is not a cognitive model for how the item is approached and solved. Especially for an item with so many components (seven constraints; five different figures to consider), one cannot assume that all examinees solve the item by mentally constructing something like a constraint satisfaction matrix, or even that all examinees solve the problem by considering the options constraint-by-constraint, successively ruling out options.[1] However, the item is written as a multiple constraint satisfaction task, and it is not unreasonable to employ the device of a constraint satisfaction matrix in representing its task structure. Moreover, and from a practical perspective, it is this representation of the task structure that facilitates the construction of an automatic item generator for variations of M161Q01.

The automatic item generator modeled on M161Q01 is referred to here as the *Triangles template*. The Triangles template randomly generates a variation on the M161Q01 stimulus text through six independent steps.

1. Select one of six different symbol sets to denote the points. The reference set is {P, Q, R, M, N, S}.

2. Select one of six logically equivalent restatements of Constraints 1 and 2, considered together.

3. Select one of two equivalent restatements of Constraint 3.

4. Select one of two distinct options for Constraint 6.

5. Select one of two equivalent restatements of Constraint 7.

6. Select one of six equivalent and logical orderings of presenting Constraints 3 through 7.

Variations on the item increase rapidly with manipulation of the answer choices. Each of the figures in the answer fields of the Triangles template is automatically sketched independently, with potential variations in rotation and reflection along perpendicular axes. (Reflections and rotations on the plane do not alter any of the properties and relations codified in the constraint satisfaction matrix.) Each figure also has at least two different options for conforming—or not, depending on its constraint satisfaction matrix parameters—to Constraint 6, while still meeting its constraint satisfaction matrix parameters for Constraint 7.

Importantly, all figures are drawn such that, like M161Q01, all constraints can be assessed by visual inspection. The template ensures, for example, that it could never be the case that if one line segment has to be larger than another, they might appear to be similar in size.[2]

Finally, all answer choices are scrambled, so that the key is not always in the same position.

The Triangles template enables one to capture a picture of the generated item, which in turn can be presented as the stimulus in another (test delivery) system, with multiple choice options *Figure A*, *Figure B*, *Figure C*, *Figure D*, and *Figure E*. This is precisely what was done to insert the appropriate generated Item 2 into each of the HITs. A randomly generated instance of Item 2 is displayed as "Item 2" in the Appendix.

Study Item Sampling

To generate the items for the study, the Triangles template was run three times for each of the six settings of Constraints 1 and 2 taken together (or *Constraint 1+2*, for short). This yielded three randomly sampled items, conditional on Constraint 1+2, for a total of eighteen distinct items. The odds of generating identical items by chance were extremely low; nevertheless the author verified that this did not occur.

For definiteness, items sharing a fixed Constraint 1+2 are here referred to as belonging to the same *type*. Recall that Constraint 1+2 posits a right triangle and identifies the point where the right angle is located. Although several logically equivalent restatements of the original "Triangle PQR is a right triangle with right angle at R" can be formulated, there is good reason to believe that items based on an alternative restatement could result in items different in difficulty. For example, a respondent may readily know or recall what is a right angle, but may not know or recall what is the hypotenuse of a right triangle, and if a hypotenuse is identified in the stimulus without reference to where the right angle is located, this could make the problem more difficult for the respondent. Thus, changes in Constraint 1+2 can potentially result in

variants (items that differ in difficulty) rather than *isomorphs* (items that have identical psychometric characteristics).

The author investigated six alternative restatements of Constraint 1+2. These are described and labeled in Table 2.

Table 2

Variations on Constraint 1+2

Statement of Constraint 1+2 in the Reference Symbol Set	Label for Type
PQR is a right triangle with right angle at R.	Standard
The lines PR and RQ are perpendicular to each other and form triangle PQR.	Perpendicular
The line PQ is the hypotenuse of right triangle PQR.	Hypotenuse
The line PQ is across from the right angle of triangle PQR.	Across
The longest side of right triangle PQR is PQ.	Longest
Right triangle PQR has acute angles at points P and Q.	Acute

These particular alternative restatements are of interest because, for tests such as PISA, which cover quite broad constructs, inferences based on test scores often imply that relatively minor variations in problem presentation should not affect those inferences. The way in which a person's level of mathematical literacy draws upon his or her knowledge of the relationships between parts of a right triangle should not depend on variations that are often taught and learned together. In other words, one would expect those inferences to hold regardless of these variations.

To ensure greater variety within the types, the author further checked that there was at least one item with the "inside" variation of Constraint 6, and at least one with the "outside" variation. A fourth or subsequent item was generated to replace the third if necessary, until this condition was met.

Data Collection

PISA is administered to a sample of all 15-year-olds enrolled in schools, in countries or economies enrolled in PISA (OECD, 2012). The aim of the cognitive PISA instrument in mathematics is to assess the *mathematical literacy* of students in a country, currently defined as

> [An] individual's capacity to formulate, employ, and interpret mathematics in a variety of contexts. It includes reasoning mathematically and using mathematical concepts, procedures, facts, and tools to describe, explain, and predict phenomena. It assists individuals to recognise the role that mathematics plays in the world and to make the well-founded judgments and decisions needed by constructive, engaged and reflective citizens. (OECD, 2010, p. 4)

Although OECD makes no claims as to the appropriateness of the PISA instrument for populations other than 15-year-olds, one can apply the same or a similar concept of mathematical literacy to adult populations, since the concept is considered part of what constitutes being a "constructive, engaged, and reflective" citizen.

This is the basis for administering the study items to an adult population, for the proof-of-concept purpose this study. The items were administered to workers on the Amazon Mechanical Turk (Amazon, 2012) system in the form of human intelligence tasks (or HITs), with a five-item test form constituting a HIT.

The HITs were created so that equal numbers of responses would be obtained on each of the eighteen variations on the study item (Item 2). A target of thirty responses for each variation was set. This resulted in the creation of 540 HITs.

Several measures were taken to ensure the quality of responses. One of these was to permit only workers who had a high HIT approval rating—a default setting on the Mechanical Turk system—to view and accept a HIT. Workers were informed that the HIT consisted in taking a short quiz assessing geometry knowledge as part of a research study. All workers on the system were paid upon approval of their work on a HIT; compensation was set for the study HITs, but to

discourage random responses, workers were informed prior to accepting a HIT that they would be paid only if they answered three out of five questions correctly. Workers had the opportunity to view a HIT before accepting it, a source of potential self-selection, addressed later. (On the Mechanical Turk system, it is not possible directly to disallow a worker's viewing a HIT before accepting it.)

A total of 297 valid cases were collected. A case was considered valid if it represented a unique worker IDs first attempt at any HIT, regardless of the number of correct responses. This of course fell short of the target of 540, but a reasonable range of cases was obtained across the eighteen variations of HITs (and hence the eighteen variations on the study item). For traditional item analysis, a substantial number of cases should be collected. Although the threshold is lower for a proof of concept analysis, it is especially important to consider results in light of the small sample size.

Results

Random Equivalence of HIT Groups

To compare the performance of each item type, it is essential that all HIT groups be considered equivalent. Each item in this study was scored as correct / incorrect, with one point for each correct response and zero for each incorrect response. The range of raw score *R* obtained from the sample of 297 responses was zero to five. Reported here are statistics on both *R* and (to assess the random equivalence of the HIT groups) *R′*, the raw score with the study item removed. Counts, means, and sample standard deviations for *R* and *R′* are shown in Table 3.

Even though the by-form counts are small for some forms, Table 3 shows that *R′* means do not vary greatly across forms, and (as expected) even less across form types.

Table 3

Sample Means and Standard Deviations for R and R′ by Form, Form Type, and Across Forms

Form	*n*	*R*		*R′*	
		M	*SD*	*M*	*SD*
			Type *Standard*		
1	14	2.64	1.44	2.07	1.16
2	14	3.43	0.98	2.57	0.90
3	13	2.77	1.48	2.38	1.21
1, 2, and 3	41	2.95	1.32	2.34	1.10
			Type *Perpendicular*		
4	22	3.36	1.49	2.73	1.39
5	20	2.90	1.37	2.40	1.11
6	14	3.86	0.64	2.86	0.64
4, 5, and 6	56	3.37	1.23	2.66	1.09
			Type *Hypotenuse*		
7	19	3.53	1.23	3.05	1.15
8	18	3.17	1.67	2.56	1.38
9	18	3.56	1.30	2.89	1.05
7, 8, and 9	55	3.42	1.42	2.83	1.20
			Type *Across*		
10	15	3.20	1.33	2.60	1.14
11	15	3.33	1.19	2.60	1.02
12	16	3.31	1.61	2.81	1.29
10, 11, and 12	46	3.28	1.39	2.67	1.15
			Type *Longest*		
13	13	3.62	1.21	2.69	1.14
14	15	3.47	1.31	2.73	1.12
15	11	3.64	1.15	2.82	1.11
13, 14, and 15	39	3.57	1.23	2.75	1.12
			Type *Acute*		
16	24	3.33	1.60	2.63	1.35
17	19	3.42	1.23	2.84	1.09
18	17	3.24	1.48	2.53	1.29
16, 17, and 18	60	3.33	1.44	2.67	1.25
1 through 18	297	3.32	1.34	2.65	1.15

Note: Sampling weights were employed in obtaining weights by form type and across all forms.

It is not possible to guarantee that no self-selection by HIT workers occurred on the basis of a specific variable feature of the study item. However, it is difficult to imagine workers engaging in such self-selection, since on the surface all versions of the study item appear the same. To investigate the possibility of self-selection, statistical analysis of performance results from other items (that is, on *R′*) was conducted. Evidence for self-selection would surface as systematic differences in *R′* means for the HIT groups.

A one-way ANOVA on the means of *R′* by form yielded no basis for rejecting the null of equivalent groups, $F(17, 279) = 0.56$ ($p = 0.921$). The design of this study is in fact one in which variations on the study item are nested within those types. To account for this structure, it was informative to re-run an ANOVA on *R′* with variation nested in type, to estimate the effect of the latter. Once more, no basis exists for rejecting the null of equal means, this time at the level of item type, $F(5, 12) = 1.91$, $p = 0.167$.

These results, together with the other measures taken to ensure equivalent groups (such as random assignment), support the assumption that the respondents in each HIT group can be considered random draws from a common population. This in turn supports the appropriateness of treating the six type-defined HIT groups as equivalent with respect to the studied item.

Common Construct Assumption

For comparisons of the study item on characteristics other than item difficulty, it is important that the non-study items and the study items all measure the same construct.

The assumption of a common construct across all 22 items (18 variable + 4 fixed) in this study is supported by positive correlation coefficients between the item scores. It is straightforward to verify this for the four fixed items (1, 3, 4, and 5). There is much reduced power in assessing item correlations for the full form, however, which includes Item 2 (the study

item). This is because very few cases (between 11 and 24) are available for each instance of Item 2.

The problem is exacerbated by the fact that not one of the 14 respondents taking Form 6 got the study item incorrect. Thus, no meaningful conclusion can be drawn about the relationship between Variation 6 of the study item and any other item. Despite this limitation, it is useful to follow up on the four-item analysis with calculations of spuriousness-corrected point-biserial correlation coefficients for each variation of the study item.

Characteristics of the fixed four-item form. If Item 2 is ignored, statistics can be computed on the remaining four-item form. The mean and standard deviation of R' on this form were 2.66 and 1.19, respectively, for the 297 respondents. The form has an estimated Cronbach's alpha reliability coefficient of 0.548 with a 95% confidence interval of (0.458, 0.627), indicating—for a four-item test—a good level of construct coherence among the items. Item statistics, presented in Table 4, show item difficulties typical of tests well aligned to the assessed population and item-total correlations typical of items assessing the same construct.

Table 4

Item Difficulties and Point-Biserial Correlations for the Fixed Four-Item Form (n = 297)

Item	P-Value	Point-Biserial
1	0.576	0.352
3	0.869	0.315
4	0.630	0.346
5	0.586	0.339

Note: The point-biserial correlation coefficients have been corrected for spuriousness—that is, the item in question was not used in calculating the total score component.

Characteristics of each five-item form. Table 5 shows the study item difficulties and their spuriousness-corrected point-biserial correlation coefficients. Due to small sample sizes, statistics of the each of the 18 five-item forms should be interpreted with caution.

Table 5

Item Statistics for the Study Item for each Five-Item Form

Form	*n*	P-Value	Point-Biserial[a]
		Type *Standard*	
1	14	**0.57**	*0.43*
2	14	*0.86*	**0.03**
3	13	**0.38**	*0.40*
		Type *Perpendicular*	
4	22	0.64	**0.06**
5	20	**0.50**	0.36
6	14	*1.00*	–[b]
		Type *Hypotenuse*	
7	19	**0.47**	**-0.04**
8	18	0.61	*0.49*
9	18	0.67	0.37
		Type *Across*	
10	15	0.60	0.19
11	15	*0.73*[c]	0.21
12	16	**0.50**	*0.53*
		Type *Longest*	
13	13	*0.92*	0.18
14	15	*0.73*[c]	0.26
15	11	*0.82*	**-0.08**
		Type *Acute*	
16	24	0.71	*0.43*
17	19	0.58	**0.07**
18	17	0.71	0.27

Note: (a) Corrected for spuriousness; (b) could not be estimated from the data. The five lowest p-values and point-biserials are in bold; the five highest are italicized. (c) These two values are tied, effectively resulting in six (not five) highest p-values.

As expected, the range of estimated item difficulties is large, as is the range of item-total correlation coefficients. All but two of the latter are positive (not considering Form 6). The two negative point-biserials are only slightly below zero. This may be due to chance variation among the 18 estimates, each based on small samples.

Two special observations are made here. First, there is no discernible pattern of association between point-biserials and form type, further supporting the claim that all item types assess the same construct as the fixed items. Second, Type Longest emerges as consistently easier than the other types—all three of its item p-value estimates are among the largest.

Exploratory Analysis of Instrument Comparability Across Types of the Study Item

Table 6 shows the study item p-values and point-biserial correlation coefficients, for the form with the median p-value, by form type. On this comparison, Type Perpendicular proved to be the most difficult among the six median-difficulty forms, and Type Longest appeared to be the easiest. The Type Longest form with the median difficulty on the study item also showed the lowest item-to-total correlation (negative, in fact), possibly partly due to its relatively high item p-value.

Table 6

Study Item Statistics for the Form with the Median Study Item Difficulty by Form Type

Item 2 Instance	Type	P-Value	Point-Biserial
1	Standard	0.57	0.43
5	Perpendicular	0.50	0.36
8	Hypotenuse	0.61	0.49
10	Across	0.60	0.19
15	Longest	0.82	-0.08
18	Acute	0.71	0.27

Note: The point-biserial correlation coefficients have been corrected for spuriousness.

How likely is it that the variation in estimated difficulty for the study item is due to chance? The question is difficult to answer because we do not know what the difficulty of the study item would be, under the null hypothesis that each variation is equally difficult. Moreover, posing the question for every one of the 18 forms inflates the Type I error rate due to multiple comparisons. One can attempt to answer the question, however, by estimating a fair "global" item p-value for the study item, and substituting that as the p-value under the null. Moreover, one can run the statistical test at the level of form type, reducing the number of tests and also increasing power by aggregating across versions of the same type.

As for a global item difficulty, the percent correct on Item 2 (regardless of form) was used. This value is 0.66. Two-sided exact binomial tests were applied to the observed number of correct responses on the study item for every item type. The statistical p-values for the two-sided exact binomial tests range from a low of 0.04 (for Type Longest) to a high of 1.00 (for type Acute). Longest is the only item type the estimated difficulty of which is statistically significantly different from the global difficulty. (The next highest statistical p-value for these binomial tests was that for Type Hypotenuse, at $p = 0.26$.)

These results suggest that Longest is an outlier among the types. This in turn points to the original six-level Constraint 1+2 as a potentially radical feature.

Considerations Prior to Revising the Template

A second look at a particular level of a variable item feature can have three general outcomes in the framework of AIG, depending on what option is taken by a test developer. The first option is to retain the feature as is, with recognition that both isomorphs and variants are possible under the item-generating template. The benefit here is greater flexibility at the potential cost of a more complicated empirical model to account for variation in item difficulty (and possibly discrimination, pseudo-guessing, or any other estimated parameters) across the variants.

The second outcome, if the feature is set to a constant value (for example, kept in the form of the original item, as in the type labeled Standard), is that the feature is essentially removed as a variable in the item-generating template. This outcome has the benefit of simplicity. Templates that generate isomorphic items are easier to employ in generating instances of items and entire test forms with predictable properties. One drawback is a reduction in the variety of generated items.

A third option is to restrict the range of values for the radical-leaning feature, preserving some variability but keeping the item-generating template in isomorphic territory. This is the best outcome, if it can be obtained. Like the second option, it also requires a justification for removing the problematic value or values of the feature in question.

The third option was adopted in this study, to illustrate how a reverse engineering approach to AIG can simultaneously help place practical parameters around assessment constructs and also improve the process of creating useful item generators.

Revision of the Triangles Template

As can be inferred from Table 1, logical application of Constraint 1+2 to the study item answer choices has the effect of eliminating two specific choices, regardless of item instance. And as Table 2 shows, all restatements of Constraint 1+2 are logically equivalent. What is not shown is the extent to which each one of these alternatives actually requires a full understanding of the statement in order to eliminate the two choices.

Upon further review of the alternative restatements of Constraint 1+2, it was observed that all item types, except one, require an understanding and application of the (technical) terms *right [tri]angle*, *perpendicular*, *hypotenuse*, and/or *acute*, in order to employ Constraint 1+2 to reduce the field of possible answer choices down to three. Although Type Longest contains the term *right triangle*, all five of the answer choices are such that knowing that PQ is the longest segment is sufficient to eliminate the same two out of five options. The reason is that the relative length of PQ is always greater than any other length, in any figure, and across all variations for the three figures that do indeed have a right angle at R. And so, the relationship between performance on the item and the need to understand technical terminology is weakened for one type, and the item is easier. Although it cannot be proved that this is the reason for type Longest to emerge as an outlier, it is a plausible that can be tested with further studies. The explanation is taken as a working hypothesis.

This finding illustrates, first, the need for examining dependencies between fixed features in a proposed item generating template, and the knowledge requirements of different potential values of a template variable. More fundamentally, it makes clear how a reverse engineering approach to AIG must negotiate between *adhering to* the model item and *generalizing from* it.

In our case, an AIG developer has the option of introducing more variability in the answer options, introducing for example a dependency between Type Longest and how one the distracter figures is displayed, restoring the dependence of item performance on understanding and applying the targeted technical terminology. This can be done with the understanding that now the field of targeted technical terms has expanded beyond just the *right [tri]angle* of the model item.

Alternatively, the developer can adhere more closely to the original, and note that the resulting generator supports only more modest inferences regarding the component of the construct that deals with the application of technical terminology. (Perhaps only "elementary" technical terms are included, for example.)

In the interest of generalizing from the template but not creating additional dependencies between its parts, the Type Longest was removed from the field of values that Constraint 1+2 can accept. This action lowered the overall item p-value across the (now) five item types. Acknowledging the limitations of formal statistical testing is this exploratory context, re-applying the exact binomial tests to each of the types under the null hypothesis that the population value for the study item's difficulty is its revised global estimate (of 0.64), yields two-sided test *p*-values ranging from 0.405 to 0.760—very weak evidence for rejecting the null of no item type effect.

A robust estimation of the characteristics of the items generated from the revised template is beyond the scope of this study for three reasons, the last two of which were present in this study by design. First, the size of the sample is too small for obtaining sufficiently accurate item p-values and point-biserial correlation coefficients. Second, the sample is not representative of individuals for whom the test was designed. Third, the form into which the study item was

embedded is not anchored to the target domain, but created as a proxy for the limited purposes of this study. Nonetheless, the revised Triangles template, and any template having undergone a similar study, would be eligible for incorporation into a field test, with anchors from the target domain, and administered to the target population for estimation of item statistics. The design of such a trial should ideally permit the testing and estimation of statistics of items derived from more than just one template.

Discussion

In this study, a reverse engineering approach to automatic item generation (Gierl & Haladyna, 2013; Haynie, et al., 2006; Masters, 2010) was applied to a figure-based publicly released test item from the PISA mathematical literacy cognitive instrument (OECD, 2006).

The essential elements of a reverse engineering approach were identified as (a) a justified exemplar, (b) a justified item schema, and (c) an item generator. This study focused on developing an item schema and its generator (or template), and on exploring the psychometric homogeneity of items produced by that generator.

Evidence was uncovered that one level of a variable feature of the generated items led to potential systematic differences in difficulty. Reasons were postulated as to why this effect surfaced, and the template was revised to more closely generate isomorphs.

This study demonstrates the feasibility of implementing a reverse engineering approach to generate multiple equivalent versions of an item from an important testing program. The study shows that it is possible, moreover, to do so even when item creation requires the production of a set of randomly generated figural elements adhering to preset constraint satisfaction parameters.

The process outlined in this study is scalable to the level of test forms or test events, but only with significant initial investment of resources and the careful implementation of a

framework such as evidence-centered design (Mislevy, Steinberg, & Almond, 2002; see also Huff, Steinberg, & Matts, 2010; Huff, et al., 2013) or assessment engineering (Lai & Gierl, 2013; Luecht, 2013; Luecht, 2007).

Limitations

Two built-in limitations of the study are its reliance on results from a non-target population and non-target fixed-form items. These limitations are acceptable in a proof-of-concept setting, if the choice of alternative population and items are justified. Were reverse engineering to be applied to automatic generation for an operational program, a study such as this would be conducted through careful embedding of item variations in (ideally) operational conditions.

A fundamental limitation of this study, and ultimately a limitation of any similar study of AIG, is that it is not feasible to test for the equivalence of every instance generated by a template. Some assumptions must be made for the technology to be practical. One of these is the assumption that certain features are incidental and truly do generate isomorphs. For the Triangles template, these features might include the set of symbols used to designate geometric points, the particular angle of rotation of a figure, and the order in which a small set of constraints is introduced. Without being able to assume the existence of isomorph-generating features, the need to test every variation of an item could easily defeat the purpose of producing a useful template in the first place. The implications of this are critical for the feasibility of AIG.

Implications for AIG Theory and Practice

The postulation of isomorphs under automatic item generation models raises two important questions for AIG theory and practice. The first is empirical. If AIG templates contain variable elements considered to be incidental, to what extent should any observed variability in

item characteristics across diverse instances of the item be investigated and, if uncovered, be used to revise the set of admissible values for those incidentals? The answer to this question has practical implications for the viability of AIG in operational settings. As pointed out, if AIG practitioners cannot assume isomorphs, or must be open to confirming them fully in every instance, then a great deal of the benefit of AIG is subverted.

There is also a theoretical question raised by the admission of isomorphs. Testing specialists know that any change in the specific instrument of measurement poses a potential threat to the equivalence of test scores. This is the fundamental reason for their being extensive research on equating, without which the field would be unable to rely on alternate forms or make score-based inferences through a common form or underlying scale.

Admitting isomorphs entails an additional step in our validity arguments: Recognizing that (at least some of) the items in test instruments are each sampled from a domain defined by all admissible combinations of the item's variable features, perhaps even by a more complex domain consisting of alternative "content equivalent" templates, each containing their own item domains. How much of that structure should one account for in theories of generalizability and in the computation of indices of accuracy, reliability, and validity?

A reverse engineering approach to AIG confronts this question directly. It opens up the possibility that the isomorph-variant distinction is not as clean as practitioners would like it to be. The extent to which one considers the need to verify the empirical properties of each combination of features for any template will determine an important threshold for validation agendas.

Reverse engineering addresses the content-equivalence precondition of equating in a systematic way. Its analog in traditional test maintenance would be a process in which several

independent groups are tasked to create a form equivalent to one or more base forms. The cost, time, and effort of creating only one such form is enough to dissuade test sponsors from attempting such a parallel construction task. The field may set aside such an endeavor for good, practical reasons, but it remains an unexplored part of test validation.

AIG through reverse engineering reconsiders this territory. Whether parallel forms are created in a traditional manner or through templates, the measurement assumption is the same: Distributions of scores on any one of the forms is a random draw from the sample population distribution. The fact that one cannot explore this assumption for traditional test construction as fully as one might through reverse engineering should help shed light on largely unexplored dimensions of measurement.

References

Amazon (2012). Amazon Mechanical Turk. website. https://www.mturk.com/mturk/welcome

Bejar, I. I. (2002). Generative testing: from conception to implementation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 199-218). Mahwah, New Jersey: Earlbaum.

Gierl, M. J., & Haladyna, T. S. (2013). Automatic item generation: an introduction. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: theory and practice* (pp. 3-12). New York: NY: Routledge.

Haynie, K. C., Haertel, G. D., Lash, A. A., Quellmalz, E. S., & DeBarger, A. H. (2006). *Reverse engineering the NAEP floating pencil task using the PADI design system*. (PADI Technical Report No. 16). Menlo Park, CA: SRI International.

Huff, K., Alves, C. B., Pellegrino, J., & Kaliski, P. (2013). Using evidence-centered design task models in automatic item generation. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: theory and practice* (pp. 102-113). New York: NY: Routledge.

Huff, K., Steinberg, L., & Matts, T. (2010). The promises and challenges of implementing evidence-centered design in large scale assessment. *Applied Measurement in Education*, *23*(4), 310-324.

Irvine, S. H. (2002). The foundations of item generation in mass testing. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 3-34). Mahwah, NJ: Lawrence Erlbaum Associates.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

Lai, H., & Gierl, M. J. (2013). Generating items under the assessment engineering framework. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: theory and practice* (pp. 77-89). New York: NY: Routledge.

Luecht, R. M. (2007). *Assessment engineering in language testing: From data models and templates to psychometrics.* Invited paper presented at the National Council on Measurement in Education, Chicago, IL.

Luecht, R. M. (2013). An introduction to assessment engineering for automatic item generation. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: theory and practice* (pp. 59-76). New York: NY: Routledge.

Masters, J. S. (2010). *A comparison of traditional test blueprinting and item development to assessment engineering in a licensing context.* (Doctoral dissertation, University of North Carolina at Greensboro). Retrieved from http://libres.uncg.edu/ir/uncg/listing.aspx?id=3646

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the roles of task model variables in assessment design. In S. Irvine, & P. Kyllonen (Eds.), *Generating items for cognitive tests: Theory and practice.* (pp. 97-128). Hillsdale, NJ: Lawrence Erlbaum.

OECD. (2006). PISA mathematics released items. Author. Retrieved from www.oecd.org/pisa/38709418.pdf

OECD. (2010). PISA 2012 mathematics framework. Draft. Author. Retrieved from http://www.oecd.org/pisa/pisaproducts/46961598.pdf

OECD. (2012). PISA Frequently Asked Questions. Author. http://www.oecd.org/pisa/pisafaq/

Smith, J. (1982). Converging on correct answers: A peculiarity of multiple-choice items. *Journal of Educational Measurement*, *19*(3), 211-220.

Footnotes

[1]The author hereby thanks a colleague who, during a general discussion of sampling incorrect choices for the generation of random distracter set, pointed out the pitfall documented by Smith (1982), in which a test wise examinee is able to solve an item by exploiting the plausibility requirement for distracters, considering features associated with all options presented, and selecting the option for which those features converge. If an item is convergence strategy dependent, the test wise examinee's selected option will also happen to be the key. As an informal test of the convergence strategy dependence of M161Q01, the author showed that item's options to the colleague, to which the colleague applied the convergence strategy, and arrived at the answer. Interestingly, the features selected by the colleague were not the same as the features in the stimulus. Thus the colleague obtained the item correct by comparing the figures in a manner different from what was called for in the stimulus, thereby bypassing any standard task structure.

[2]It would have been possible to generate an even greater variety of figures for M161Q01, by allowing, for example, hundreds or thousands of different options for the relative lengths of line segments or sizes of angles. This level of variation is undesirable for a number of reasons, one of which is the risk that item difficulty can change greatly for construct-irrelevant reasons, such as a person's ability to distinguish between angles of 90 and 95 degrees, or to judge a line segment to be larger than another, when they are not presented side by side and the lengths differ only by a small amount.

Appendix A: Sample HIT

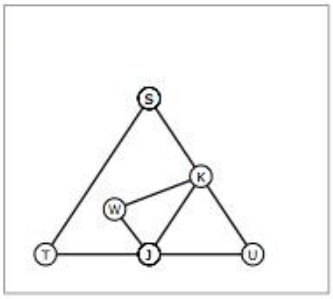
1. Which of the following is longer than the circumference of a circle?

- The perimeter of a square inscribed in the circle
- Three times the diameter of the circle
- Seven times the radius of the circle
- Nine times the arc length of the circle at 30 degrees

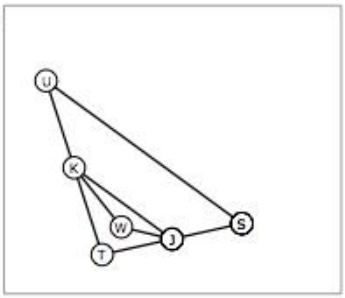
2. Which of the following figures fits the description below?

The lines UT and SU are perpendicular to each other and form triangle STU. W is a point inside the triangle. The line UT is less than the line SU. J is the midpoint of the line ST and K is the midpoint of the line TU. The line JK is greater than the line JW.

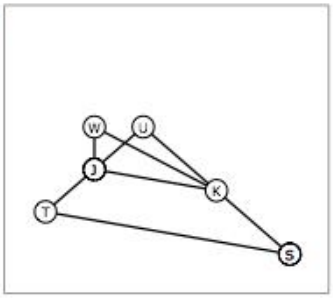
A



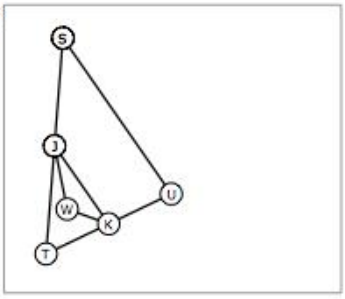
B



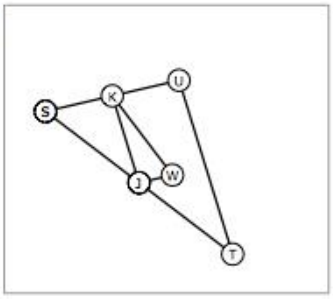
C



D

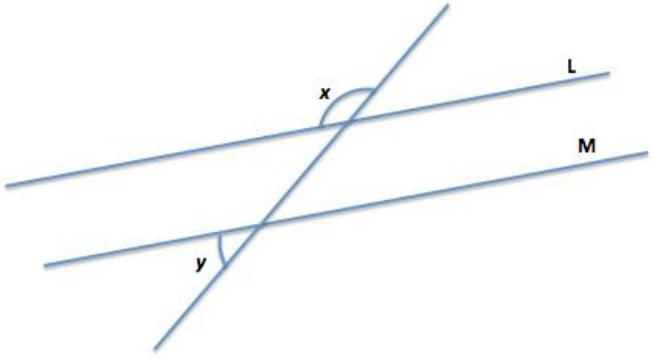


E



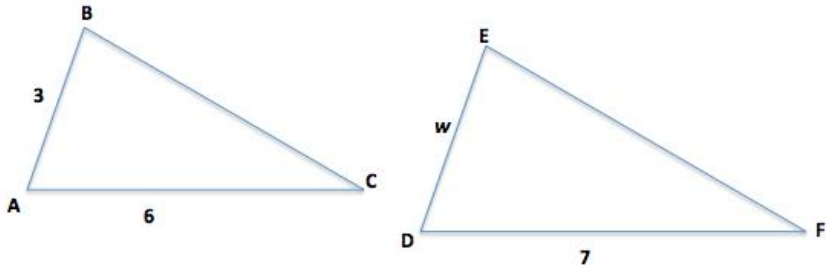
- A
- B
- C
- D
- E

3. In the figure below, lines L and M are parallel. Angles x and y are marked. True or False: $x + y = 180$ degrees.



◯ True ◯ False

4. Triangle DEF is an enlarged copy of triangle ABC. Angle A is equal to angle D and angle C is equal to angle F. What is the length of side w?



◯ 7/5

◯ 5/3

◯ 6/7

◯ 7/2

5. In a regular solid, an *edge* is a line where two surface figures meet. An octahedron (pictured below) is a regular solid with 8 surface triangles. How many edges does an octahedron have?

- 4
- 6
- 8
- 12
- 16