Chapter 6
On the Number of Items in Testing Mastery of Learning Objectives

Anton A. Béguin and J. Hendrik Straat

Abstract In individualized learning trajectories, it could be valuable to administer small tests that focus on a specific learning outcome to determine mastery of the learning objective and to evaluate whether a student can progress to other learning objectives. For this type of application, testing time competes with direct learning time, and a large number of learning objectives could invoke a potentially large burden due to testing. Thus, it is effective to limit the number of items and to reduce testing time as much as possible. However, the number of items is directly related to the accuracy of the mastery decision and the applicability of this type of formative evaluation in practical situations. For the formative evaluation to result in valid inferences, general measurement principles are valuable as well (Bennett in Assess Educ Principles Policy Pract 18:5–25, 2011). In this chapter, we provide techniques to determine the number of items and corresponding cut scores that are necessary to decide on mastery. We apply these techniques in situations with different item characteristics and provide the outcomes for varying test situations, illustrated using a practical example.

6.1 Introduction

The requirements for mastery testing and classification testing have been studied quite extensively (e.g., Wilcox 1976; de Gruijter and Hambleton 1984; van der Linden 1990; Vos 1994; Van Groen 2014). The earlier research focused on the proportion of items mastered in a well-specified content domain, containing all the relevant items in that domain (Hambleton and Novick 1973; de Gruijter and Hambleton 1984). Here, the domain is a hypothetical concept that contains all the possible items in this content domain. This proportion is referred to as π and can be interpreted as the true proportion-correct score of a person on this domain. The standard on the domain is defined as π_0, and if $\pi \geq \pi_0$, the person has mastered the domain. In practice,

A. A. Béguin (✉) · J. H. Straat
Cito, Arnhem, The Netherlands
e-mail: anton.beguin@cito.nl

B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, https://doi.org/10.1007/978-3-030-18480-3_6

only a sample of these items can be administered in a test. The score on the test, x, is evaluated against a cut-score C_x. Theoretically, C_x can be defined as $C_x = n\pi_0$, assuming equal difficulty for all items and perfect reliability, with n the number of items in the test. In reality, the assumptions are never met and misclassifications will occur if $\pi \geq \pi_0$ and $x < C_x$ or if $\pi < \pi_0$ and $x \geq C_x$. It can also be shown that the above definition leads to a non-optimal cut-score when the population mean is higher or lower than the standard, when the reliability is low, and when false positives are valued differently than false negatives (de Gruijter and Hambleton 1984). Alternative approaches to set a cut-score are based on utility theory and Bayesian decision theory (van der Linden 1980).

Wilcox (1976) reported on the appropriate lengths and passing scores based on true score proportions π_0 of 0.7, 0.75, 0.8, and 0.85. To determine the percentage of correct decisions, he defined a zone of indifference around π_0. This zone of indifference varies between $\pi_0 - 0.05$ and $\pi_0 + 0.05$, and in another condition, between $\pi_0 - 0.1$ and $\pi_0 + 0.1$. Individuals with true score probabilities within this interval will not be evaluated as incorrectly classified, independent of whether they score below or above the threshold on a test. He found that 80% or more correct classifications are established for conditions with an indifference zone of $\pi_0 \pm 0.1$ with a 19-item test, $C_x = 14$ and $\pi_0 = 0.7$. With $\pi_0 = 0.8$, this percentage is reached with an 18-item test and $C_x = 15$, while for $\pi_0 = 0.85$, an 11-item test and $C_x = 10$ is sufficient.

Other research related to mastery has focused on the application of item response theory (Rasch 1960; Lord 1980) to scale items in a test form and to determine the accuracy of person-parameter estimates. Item response theory can correct for the differences in difficulty between items and test forms that occur due to sampling from the domain. In line with this, research has been done on mastery and classification decisions, applying adaptive testing procedures (Eggen 1999; Eggen and Straetmans 2000; Van Groen 2014).

A different approach to decide on both test lengths and cut-scores of mastery tests can be based on informative Bayesian hypotheses (Hoijtink et al. 2014). Following their description, mastery is determined based on responses to a set of items. Given the person is a master, the minimum probability of answering each item correctly is determined. This leads to informative hypotheses for each of the items in the set of items. For example, if it is assumed that masters have a probability of 0.8 or more to answer an item i correctly, the following hypothesis is used:

$$H_{i,master} : \pi_i > 0.8.$$

Aggregating over all items $i = 1, \ldots, \mathrm{I}$, and assuming independent non-informative uniform prior distributions (Gelman et al. 2004) on the interval [0–1], the prior for $\pi = [\pi_1, \ldots, \pi_I]$ is:

$$h(\pi) = \prod_{i}^{I} \mathrm{Beta}(\pi_i|1, 1) = 1.$$

The posterior distribution given responses $x = [x_1, \ldots, x_I]$ is:

$$g(\pi|x) \propto h(\pi) \prod_{i}^{I} \pi_i^{x_i} (1 - \pi_i)^{1-x_i}$$

$$\propto \prod_{i}^{I} \text{Beta}(\pi_i|1 + x_i, 1 + (1 - x_i)).$$

The proportion of the posterior in agreement with the mastery hypotheses is:

$$f_i = \int_{\pi \in H_{i,master}} g(\pi|x) \partial\pi,$$

with $i = 1, \ldots, I$.

If we also determine the proportion of the prior in agreement with the mastery hypotheses:

$$c_i = \int_{\pi \in H_{i,master}} h(\pi) \partial\pi,$$

a Bayes factor (Kass and Raftery 1995) can be determined, comparing the informative mastery hypotheses (m) to hypotheses without constraints (u):

$$BF_{mu} = \frac{f_i}{c_i}$$

By the same token, hypotheses can be defined by focusing on the response behavior of non-masters. This can be the complement of the behavior of masters, thus, any response pattern not meeting the criteria for masters, or it can be a set of hypotheses with additional restrictions of their own. For example, a restriction that the probability of answering an item correctly for a non-master is smaller than 0.4 for each of the items is:

$$H_{i,non\text{-}master} : \pi_i < 0.4.$$

Obviously, all kinds of other hypotheses are possible, and also hypotheses that differ per item can be combined. For example, if a researcher adopts the diagnostic perspective as formulated by Hoijtink et al. (2014), one could use latent class analysis (LCA; Lazarsfeld 1950) to define groups of masters and non-masters. More complex constructions of classes can be considered by putting restrictions on the probabilities of answering items correctly, given class membership (e.g., Heinen 1996; Hoijtink 2001; Vermunt 2001). The Bayes factor can then be used to test the most likely class membership, given a specific score pattern.

In the current research, we will apply informative Bayesian hypotheses to evaluate test lengths and cut-scores for items typically used in mastery testing, with a focus on fine-grained learning objectives. Typically, the items in assessments that focus on mastery of a learning objective are constructed in such a way that students who have mastered the learning objective will have a high probability of answering the items correctly. Students who have not mastered the learning objective will have a smaller probability of answering the items correctly. We establish guidelines for test lengths and cut-scores in three studies: a simulation study with homogeneous item characteristics, an empirical example, and a simulation based on the empirical example with heterogeneous item characteristics.

6.2 Method

6.2.1 Simulation Study with Homogeneous Item Characteristics

We evaluated the Bayes factors for number-correct scores on tests with 4–10 items. Mastery on these tests was defined as having a probability higher than 0.8 to answer each of the items correctly. For non-mastery, four different hypotheses were considered. The first hypothesis to define non-mastery was that at least one item should have a probability of being correctly answered lower or equal to 0.8. This is the complement of the definition of mastery given above. The three other hypotheses that defined non-mastery were that the probability of giving a correct answer to an item was smaller than 0.2, 0.4, or 0.6 for all of the items. The Bayes factors for mastery compared to each of these alternatives for non-mastery were calculated using the program BED.exe (Hoijtink et al. 2014).

To interpret the Bayes factors in this study, we followed the proposed guidelines in the literature (Kass and Raftery 1995; Jeffreys 1961) and adopted the rule that Bayes factors over 20 are an indicator of mastery. According to the guidelines, these values are an indication of strong evidence (BF between 20 and 150) or very strong evidence (BF > 150) that the response is based on mastery rather than non-mastery. The rationale behind the somewhat conservative rule and not accepting lower BF values is that in formative evaluations, the cost of false negatives is relatively low, while due to a false positive decision, a student could miss extra education on a topic that needed more attention.

6.2.2 Empirical Example

We applied the Bayesian hypothesis testing method to item response data collected from *Groeimeter* (2017), an evaluation platform containing mastery tests for a large

number of mathematics learning objectives. Each learning objective is assessed by a 7-item test and a student takes multiple tests. The data of performances on 25 formative tests were used in Bayesian evaluations of mastery of learning objectives based on inequality constrained hypotheses identified through latent class analyses. Each formative test was evaluated separately using the following two steps:

Step 1. The probabilities of answering the seven items correctly were determined separately for masters and non-masters. In the data, both groups were present, since the formative tests were administered to students who were either masters or used the formative tests for practice. The specific test strategy for a single student was unknown to us; thus, we used latent class analyses (*poLCA;* Linzer and Lewis 2014) to identify the classes of students, which were then interpreted as masters and non-masters. The success probabilities for item i for masters $\pi_{i,masters}$ and non-masters $\pi_{i,non\text{-}master}$, were used to specify hypotheses in which these probabilities define the borderline case for mastery and non-mastery. This resulted in inequality constrained hypotheses $H_{i,master} : \pi_i \geq \pi_{i,masters}$, and $H_{i,non\text{-}master} : \pi_i \leq \pi_{i,non\text{-}master}$.

Step 2. Each of $2^7 = 128$ possible score patterns were evaluated against both sets of hypotheses. If the Bayes factor for mastery against non-mastery exceeded 20, it was concluded that the response pattern corresponded to a student who had mastered the objective. For each learning objective, the Bayes factors were calculated using the program BED.exe (Hoijtink et al. 2014). Subsequently, score patterns resulting in a Bayes factor of 20 or higher were classified as indication for mastery. Since all items differ in the probabilities for mastery and non-mastery the specific score pattern impacted the Bayes factor. Patterns with equal number-correct score but a different score pattern could lead to a different indication for mastery. The minimum number-correct score for mastery was determined based on the proportion of patterns with the same number-correct score leading to a mastery decision.

6.2.3 Simulation Study Based on Empirical Data and Heterogeneous Item Characteristics

The empirical example used the results of 25 * 7 = 175 separate items from 25 different learning objectives. These items psychometrically reflected a wide range of item characteristics that can be found in real data. The relevant item characteristics were the success probabilities for masters and non-masters from the latent class analyses. These probabilities were used to define the inequality constraints for mastery and non-mastery as described in step 1 above. Based on the set of 175 items, new formative tests were assembled with different test lengths.

The required number-correct score for mastery was determined for tests with 4–10 items. For each test length, we simulated 50 replications by drawing from the 175 items without any replacements. We then estimated the Bayes factor for all the possible response patterns for inequality constrained hypotheses for masters

and non-masters (similar to Step 2 in the analyses of the original empirical data). This was done to evaluate the effectiveness of different test lengths and different number-correct scores to distinguish between masters and non-masters.

6.2.4 *Estimating and Validating a Predictive Model for Bayes Factors*

To aggregate the results over the different tests from Groeimeter, a regression model was estimated in which the Bayes factor was predicted based on the response pattern and taking into account item characteristics. Aggregation was necessary since tests for different learning objectives will show variations in item characteristics and consequently in the required number of correct responses to indicate mastery. The dependent variable was the natural logarithm of the Bayes factor, accounting for the non-linear nature of this variable. Four predictors were used: (1) an intercept, (2) the observed proportion correct of the response pattern, (3) the sum of the success probabilities for masters on the incorrect responses, (4) the sum of the success probabilities for non-masters on the correct responses. The last two variables were centralized around the mid-point of the probability scale.

Results from the analysis based on the data from *Groeimeter* were validated with results calculated on the generated samples from the simulation study.

6.3 Results

6.3.1 *Simulation Study with Homogeneous Item Characteristics*

Results of the simulation study that focused on the number-correct score and test length are given in Table 6.1. The four conditions are indicated in the first column and are a single definition of mastery, with π larger than 0.8 and indicated by (m: > 0.8), crossed with each of the four conditions of non-mastery, ranging from the complement of all π larger than 0.8 (> 0.8) down to all $\pi < 0.2$. Within each condition, Bayes factors are given for test lengths of 4–10 items. Bayes factors 20 or higher are printed in italics. For each test length n, all of the possible number-correct scores 0 … n were evaluated, but only a limited number of results are reported. Indications of non-mastery and very large Bayes factors are removed from Table 6.1. This includes all factors smaller than 0.2 and larger than 1000.

The Bayes factors in Table 6.1 can be evaluated to find appropriate test lengths and cut-scores for mastery. For example, it can be seen that no Bayes factor was larger than 20 for the 4-item and 5-item tests in condition 1. For tests with lengths of 6–8 items, only a perfect score indicates mastery in condition 1, while a number-correct

Table 6.1 Bayes factor comparing mastery and non-mastery

Condition	Number correct	4	5	6	7	8	9	10
(1) m: > 0.8 nm: > 0.8	n	10.9	19.5	*35.3*	*63.7*	*111.3*	*195.0*	*372.4*
	$n-1$	1.2	2.1	3.9	7.0	12.5	*22.0*	*41.0*
	$n-2$		0.2	0.4	0.8	1.4	2.4	4.6
	$n-3$						0.3	0.5
(2) m: > 0.8 nm: < 0.6	n	*84.6*	*254.3*	*777.4*				
	$n-1$	4.0	12.0	*36.8*	*109.7*	*326.0*	*944.0*	
	$n-2$	0.2	0.6	1.7	5.2	15.6	*45.0*	*142.4*
	$n-3$				0.2	0.7	2.1	6.7
(3) m: > 0.8 nm: < 0.4	n	429.7						
	$n-1$	11.7	*53.2*	*247.4*				
	$n-2$	0.3	1.5	6.8	*30.2*	*134.9*	*577.0*	
	$n-3$				0.8	3.8	16.2	*76.8*
	$n-4$						0.4	2.1
(4) m: > 0.8 nm: < 0.2	n							
	$n-1$	*81.6*	*736.9*					
	$n-2$	1.0	9.0	*84.5*	*742.6*			
	$n-3$			1.0	9.1	*90.4*	*727.8*	
	$n-4$					1.1	9.0	*80.9*
	$n-5$							1.0

score of 8 is also a clear indication of mastery in a 9-item test, and a number-correct score of 9 indicates mastery in a 10-item test.

6.3.2 Empirical Example

Subsequently, the results of the latent class analyses are given to determine success probabilities for masters and non-masters and the resulting Bayes factors for the 25 formative tests.

6.3.2.1 Latent Class Analyses

Figure 6.1 summarizes the results of the latent class analyses for the 25 formative tests sampled from *Groeimeter*. Each plot shows the distributions of $25 * 7 = 175$ different items. The three distributions represent (a) the latent class-based estimated success probabilities for the masters, (b) the estimated success probabilities for the non-

masters, and (c) the difference between those success probabilities for the masters and the non-masters.

On average, masters had a success probability of 0.84 on the test items, whereas the non-masters had an average success probability of 0.33. The probabilities for masters are close to the generally accepted boundary of 80% correct for mastery, and the success probabilities for the non-masters are low enough to enable a clear distinction between the two groups. The right panel of Fig. 6.1 shows that the difference in success probabilities differs largely across the items; one item even has a higher success probability for the non-masters than for the masters. This is a suitable collection of items to investigate the impact of differences in success probabilities on the resulting Bayes factor.

6.3.2.2 Bayes Factors

We investigated the Bayes factors for all possible response patterns on the seven items for each of the 25 formative tests in *Groeimeter*. We found that no response pattern with zero, one, or two correct responses showed enough evidence for mastery; six and seven correct responses were always congruent with a mastery response pattern. For the other number-correct scores, the cumulative distribution of natural logarithms of the obtained Bayes factors are given in Fig. 6.2. The cut-score to indicate a mastery response pattern on the natural logarithm scale of the Bayes factor is $\ln(20) = 2.996$.

In Fig. 6.2, the distributions for larger number-correct scores shift to the right, indicating that the Bayes factor generally increases with a larger number-correct score. For number-correct scores of 3–6, the percentage of the response patterns congruent with mastery of the learning objective was 2, 35, 91, and 99%, respectively.

To illustrate what conditions lead to more deviant conclusions, Table 6.2 shows two examples of response patterns with corresponding success probabilities for masters and non-masters. Test #3 has incorrect responses for easy items for the mastering group, resulting in a response pattern of five correct items and showing no significant evidence of mastery. In test #40, a response pattern resulting in three correct items was a clear indication of mastery when the correctly answered items had a very small success probability for non-masters (< 0.02).

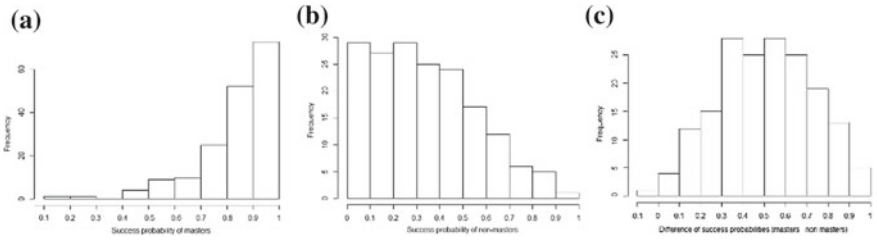

Fig. 6.1 Distributions for latent class-based success probabilities for masters and non-masters, and the difference between these probabilities

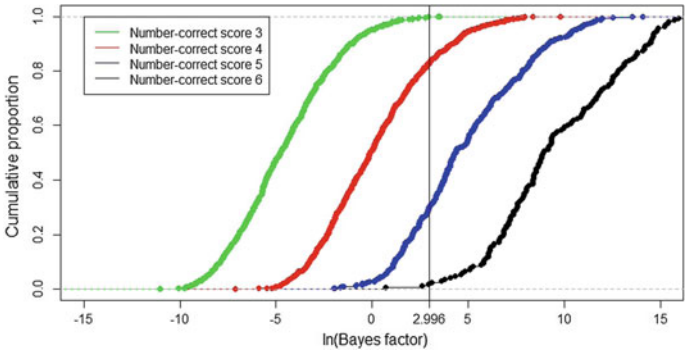

Fig. 6.2 Distribution of the natural logarithm of Bayes factors for number-correct scores 3–6

Table 6.2 Examples of response patterns leading to deviant conclusions

Test 3	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Bayes factor
Success probabilities for masters	>0.980	0.769	0.966	>0.980	0.745	0.574	0.414	
Success probabilities for non-masters	0.400	0.259	0.680	0.448	0.648	0.401	0.378	
Response pattern	0	1	1	0	1	1	1	0.141
Test 40								
Success probabilities for masters	0.529	0.973	0.850	0.895	0.822	0.868	0.719	
Success probabilities for non-masters	0.215	0.246	0.142	<0.020	0.256	<0.020	<0.020	
Response pattern	0	0	0	1	0	1	1	33.821

Table 6.3 Percentage of response patterns congruent with mastering the learning objective for different test lengths and number-correct scores

	Number-correct										
	0	1	2	3	4	5	6	7	8	9	10
4	0%	2%	11%	90%	100%						
5	0%	0%	0%	23%	95%	100%					
6	0%	0%	0%	5%	67%	97%	100%				
7	0%	0%	0%	2%	23%	89%	100%	100%			
8	0%	0%	0%	4%	9%	60%	97%	99%	100%		
9	0%	0%	0%	0%	10%	20%	81%	100%	100%	100%	
10	0%	0%	0%	0%	2%	11%	53%	96%	99%	100%	100%

6.3.3 *Simulation Based on the Empirical Data and with Heterogeneous Item Characteristics*

Tests were assembled with test lengths ranging from 4 to 10 items, and percentages of response patterns congruent with mastery were calculated for each number-correct score separately. These percentages are given in Table 6.3.

6.3.4 *Prediction Model*

The estimated regression coefficients for predicting the natural logarithm of the Bayes factors using response patterns and item characteristics are given in Table 6.4.

The effect of the proportion of correct responses can be interpreted as a modification of the intercepts, given a specific number-correct score. The larger the number-correct score, the higher the intercept. The other effects are related to the specific particular response pattern. Generally speaking, high success probabilities on correct responses for non-masters and high success probabilities on incorrect responses for masters resulted in lower Bayes factors.

Table 6.4 Regression coefficients predicting *ln*(*Bayes factor*)

Intercept	−7.112
Proportion of correct responses	15.834
Sum of success probabilities of non-masters for correct responses minus 0.5	−3.890
Sum of success probabilities of masters for incorrect responses minus 0.5	−5.229
R^2	0.953

All coefficients are significant $p < .001$

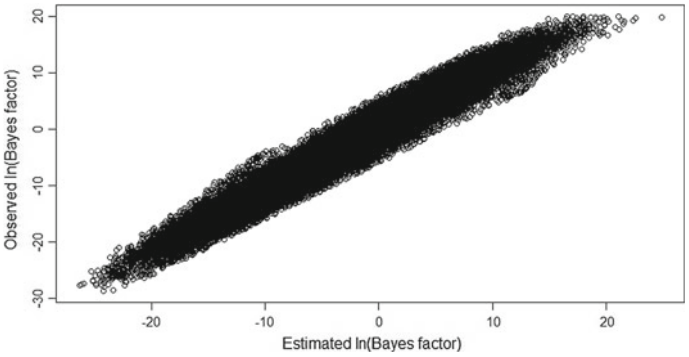

Fig. 6.3 Relationship between observed ln(Bayes factor) and the predicted ln(Bayes factor) by the regression model

6.3.4.1 Validation of the Prediction Model

The application of the regression model, as presented in Table 6.4, to all newly assembled tests in this simulation study (450 tests in total) resulted in a strong correlation ($r = 0.975$) between observed and predicted Bayes factors.

The relationship for each of the simulated tests is graphically presented in Fig. 6.3. The specificity and sensitivity of classifying response patterns as congruent or conflicting with learning objective mastery are 0.974 and 0.834, respectively.

6.4 Discussion and Conclusions

Bayesian hypotheses tests were used in a number of scenarios to answer the question: "How many correct responses do I need to decide on mastery?" As with all simulation studies and case studies, the results depend on the specific conditions, but some overall trends can be seen in the different studies. In the first theoretical study, inequality hypotheses were compared with equal difference in probability between mastery and non-mastery for all items. The amount of difference varied across conditions and in only one of the conditions the definition of non-mastery was the complement of mastery. In all other cases a comparison was made between two inequality hypotheses that did not cover all theoretically possible outcomes leaving some success probabilities unspecified as indicative for mastery or non-mastery. Probabilities between the upper bound for non-mastery and below the lower bound for mastery could provide alternative hypotheses to predict the data and be better suitable for some response patterns.

In the empirical example, our procedure incorporated results from LCA into inequality constrained hypotheses. The resulting definitions of mastery and non-mastery differed largely in success probabilities. The hypothesis tests based on these

success probabilities were extremely powerful in detecting whether or not a response pattern was in line with mastery or non-mastery. In the second simulation study even a test length of just four items provided a significant indication for mastery in 90% of the cases where a student gave three correct answers. This amount of power can be explained by two aspects:

- The LCA indicated a large difference in average success probability for masters and non-masters. This average difference was more than 0.50.
- The success probabilities are used to define two inequality constrained hypotheses that are compared, and all other hypotheses are ignored. The success probabilities are used as lower bound for mastery and as upper bound for non-mastery. Probabilities lower than the lower bound for mastery but higher than the upper bound for non-mastery were not considered as alternative to the mastery and non-mastery hypothesis, while in practice these could give alternative, and potentially more likely, explanations for the response behavior.

As a consequence the items got almost deterministic properties in the second simulation study. If an item was answered incorrectly while the probability of a correct response for masters was very high this probably resulted in a classification as non-master. By the same token, a correct answer on an item with a very low probability for non-masters probably resulted in a classification as master.

In future research, other ways to translate results from LCA into Bayesian hypotheses should be considered. For example, definitions of mastery and non-mastery could be based on mutual exclusive categories (comparable to condition 1 in the first simulation study) or an alternative procedure could be applied in which equality constraints are used to define mastery and non-mastery. Other alternatives are to use inequality constraints on the success probability plus or minus two standard errors for non-mastery and mastery, respectively, and to consider other hypotheses such as indifference about mastery or hypotheses related to specific misconceptions.

The number of items necessary to determine mastery in a test clearly depended on the conditions, the level of certainty of the mastery decision, and the cut-score used. When using a level of certainty of 95%, the difference between heterogeneous item characteristics in the second simulation study and homogeneous item characteristics in condition 3 of the first study did not result in very different outcomes. Both studies indicated mastery for a maximum score on a four item test. With tests containing five and six items a score one point below the maximum was an indication of mastery. The same was found in the heterogenous case for a test with seven items, while a score of five on a seven items test was sufficient in the homogeneous case.

When we want to allow for an incorrect item response, based on the study with homogenous inequality constraints, we need only five items when the definition of non-mastery is based on a success probability for all items of 0.4 or less. Six items is the minimum test length with a non-mastery definition based on a probability of 0.6 or less. When non-mastery is defined as the complement of mastery, at least a 9-item test with eight correct responses is necessary to indicate mastery based on a Bayes factor of 20 or more.

As a general rule, it is reasonable to assume that you at least need six items and a cut-score of 5 to be able to decide on mastery if the test is sufficiently carefully designed to perform as a mastery test. Even in that case, it is necessary to check if all items discriminate between masters and non-masters. If the items are pre-tested and all items are selected to be in line with a mastery decision for difficulty level and discrimination, the test length can be reduced to five items.

As a more general conclusion, this research showed that the evaluation of Bayesian hypotheses can provide practical guidelines for test construction and the evaluation of empirical tests. Extending on the current analyses and in line with the tradition of earlier research into mastery testing, a next step could be to incorporate utility theory (van der Linden 1990; Vos 1994) into the procedure. This can be accomplished using differential weighting of false positive and false negative decisions. Another line of research is to extend the application of the described procedure on response data of formative assessments by identifying classes that indicate particular misconceptions, thereby providing relevant feedback, given a series of responses in the case of non-mastery.

References

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18,* 5–25.

de Gruijter, D. N. M., & Hambleton, R. K. (1984). On problems encountered using decision theory to set cutoff scores. *Applied Psychological Measurement, 8,* 1–8.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with sequential probability ratio tests. *Applied Psychological Measurement, 23,* 249–261.

Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement, 60,* 713–734.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Hambleton, R. K., & Novick, M. R. (1973). Towards an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement, 10,* 159–170.

Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences*. Thousand Oaks CA: Sage.

Hoijtink, H. (2001). Confirmatory latent class analysis: Model selection using Bayes factors and (pseudo) likelihood ratio statistics. *Multivariate Behavioral Research, 36,* 563–588.

Hoijtink, H., Beland, S., & Vermeulen, J. (2014). Cognitive diagnostic assessment via Bayesian evaluation of informative diagnostic hypotheses. *Psychological Methods, 19,* 21–38.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, England: Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 773–795. https://doi.org/10.1080/01621459.1995.10476572.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis & The interpretation and mathematical foundation of latent structure analysis. In S. A. Stouffer et al. (eds.), *Measurement and Prediction* (pp. 362–472). Princeton, NJ: Princeton University Press.

Linzer, D., & Lewis, J. (2014). *poLCA*: Latent class analysis and latent class regression models for polytomous outcome variables. R package version 1.4.1.

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.
van der Linden, W. J. (1980). Decision models for use with criterion-referenced tests. *Applied Psychological Measurement, 4,* 469–492.
van der Linden, W. J. (1990). Applications of decision theory to test-based decision making. In R. K. Hambleton & J. N. Zaal (Eds.), *Advances in educational and psychological measurement* (pp. 129–156). Boston, MA: Kluwer-Nijhof.
Van Groen, M. M. (2014). *Adaptive testing for making unidimensional and multidimensional classification decisions.*
Vermunt, J. K. (2001). The use of restricted latent class models for defining and testing nonparametric and parametric item response theory models. *Applied Psychological Measurement,*
Vos, H. J. (1994). *Simultaneous optimization of test-based decisions in education* (Doctoral dissertation). Enschede: University of Twente.
Wilcox, R. R. (1976). A note on the length and passing score of a mastery tests. *Journal of Educational Statistics, 1,* 359–364.