Chapter 10

Influences on Classification Accuracy of Exam Sets: An Example from Vocational Education and Training

Marianne Hubregtse and Theo J.H.M. Eggen

Abstract Classification accuracy of single exams is well studied in the educational measurement literature. However, when making important decisions, such as certification decisions, one usually uses several exams: an exam set. This chapter elaborates on classification accuracy of exam sets. This is influenced by the shape of the ability distribution, the height of the standards, and the possibility for compensation. This is studied using an example from vocational education and training (VET). The classification accuracy for an exam set is computed using item response theory (IRT) simulation. Classification accuracy is high when all exams from an exam set have equal and standardized ability distributions. Furthermore, exams where few or no students pass or fail increase classification accuracy. Finally, allowing compensation increases classification accuracy.

Keywords: classification accuracy, misclassification, sets of exams, certification decisions

Introduction

Everyone agrees that high-stakes exams should be of sufficient quality. Quality of exams is usually studied in terms of validity and reliability for traditional standardized tests. The quality of exams that include performance assessments is generally studied in terms of authenticity (e.g. Gulikers, 2006) and the validity of the assessment (e.g. Linn, Baker, & Dunbar, 1991). In contrast to traditional forms of assessment, performance assessments are not standardized tests. This implies that traditional reliability indices may not be suitable to apply to performance assessments (Clauser, 2000; Dochy, 2009). Kane (1996) points out that the precision of measurements is broader than just reliability. Classification accuracy provides an opportunity to quantify the quality of exams in a universal way, for both standardized tests and performance assessments. Especially in cases where reliability is difficult to compute, classification accuracy may give a quantitative measure of the quality of a certain exam.

Classification accuracy is a measure of precision that may be more appropriate for performance assessments, yet also applicable to standardized tests. Classification accuracy is the degree of overlap between a decision based on the scores of an exam and the decision that would have been made on the basis of scores without any measurement error (Hambleton & Novick, 1973).

There is no measurement, and thus no exam, without measurement error. Therefore, misclassifications occur wherever the decision based on the scores of an exam deviates from a decision based on error-free scores. There are two types of misclassifications: false positives or false negatives. False positives occur when students have a true classification below the set standard, but they receive an exam score above the standard. The reverse is false negatives: students that have a true classification above the set standard fail the exam. In all other cases—true classification above the standard passes the exam and true classification below the standard fails the exam—there is no misclassification. It must be noted that classification accuracy is not the same as classification consistency. Where classification accuracy compares the true classification of a student with the observed classification, classification consistency compares the classification in two (parallel) exams (see also Lee, 2008). Here, the interest is solely in classification accuracy.

In this chapter, the words *standardized test*, *performance assessment*, *exam*, and *exam set* all have a distinct meaning. A *standardized test* is any test that includes only questions, though they can be of any type: open-ended, multiple choice, and so on. A multiple choice test is a form of a standardized test. A *performance assessment* is defined as a form of testing in which the student is asked to perform a task that she could encounter in real life. Vocational education uses performance assessments to allow a student to display competency in their prospective jobs. The word *exam* is used whenever there is need for a general word for both performance assessments and standardized tests. The term *exam set* is used for any combination and number of exams. For instance, suppose a student must do three multiple choice tests and four performance assessments in order to receive a diploma. The seven exams together would be considered one exam set.

In essence, certification decisions and other high stakes exams tend to be dichotomous decisions. Either the student scores above or below the standard set for certification. One uses an exam to classify students above or below a set threshold, in order to hand out diplomas. It is very common to base diploma decisions on more than one exam.

Usually an exam set is used that measures, as far as possible, all competencies and abilities involved. Generally, this seems like a good practice: more opportunities to observe a student give a better idea of true ability or competence. Furthermore, more measurements increase the reliability of the total measurement. Since the most important decision, the certification decision, is based on an exam set, it seems appropriate to compute the classification accuracy of that exam set instead of the classification accuracies of all separate exams.

Computing the classification accuracy of an exam set is not much different from computing the classification accuracy for a single exam. This chapter will use a method of computing classification accuracy for exam sets using simulation based on item response theory (IRT). The focus will be on how to increase classification accuracy of exam sets. In order to answer this question, three known influences on classification accuracy for single exams are studied: the shape of the ability distribution of the target population, the standards set for the exam set, and the compensation rules between the exams in the set.

Regarding the ability distribution of single exams, normalizing helps the simulation to more accurately estimate classification accuracy (Van Rijn, Béguin, & Verstralen, 2009). An increase in classification accuracy is related to more symmetrical misclassifications (Holden & Kelley, 2008), which is exactly what occurs when ability distributions are normalized.

With respect to the influence of standards on single exams, the following observations are made. Obviously, standards that are so extreme that they fall completely outside the population distribution lead to a classification accuracy of 100% (Lee, 2008). Furthermore, it is more difficult to correctly categorize students close to the standard than students far away from the standard (Martineau, 2007). Therefore, a graph of classification accuracy should show a dip around where the standard meets the average ability in the population (Lee, 2008).

Finally, the compensation rules are shown to have an effect on classification accuracy for exam sets (Van Rijn et al., 2009; Verstralen, 2009a). Both Van Rijn et al. (2009) and Verstralen (2009a) conclude that classification accuracy is increased when allowing some form of compensation. It is expected that this result is found for all other varied influences. Intuitively this makes sense, since allowing compensation essentially lengthens the separate exams into one long exam (Gatti & Buckendahl, 2006).

These three influences are studied in a simulation, using empirical data from a vocational education and training (VET) exam set as a starting point. Outcomes are expected not to differ from influences on single exams. In the next section of this chapter, the method of computing classification accuracy for the exam set is shown in short. Furthermore, the influences varied in the simulation study are discussed in more detail. The example section discusses the data source, the specific setup of the simulation study, and the results of this specific simulation. The following discussion section shows what the results imply for practitioners and discusses limitations of the study and future research.

Method

There are a few different ways of computing classification accuracy. Hambleton and Novick (1973) describe how to compute classification accuracy using two administrations of the same test. Swaminathan, Hambleton, and Algina (1974) gave a correction to this coefficient. In 1990, Huynh introduced a measure for class consistency for dichotomous items based on the Rasch model. Livingston and Lewis (1995) further elaborated on this measure, allowing for different scoring procedures. Schulz, Kolen, and Nicewander (1999) introduced the first real IRT model for estimating classification accuracy, though still only for dichotomous items. Wang, Kolen, and Harris (2000) extended this to a procedure for computing classification accuracy with polytomous IRT models. Verstralen (2009b) developed this into a method for computing classification accuracy for exam sets. This last method is used for the simulation and is explained further in the following paragraph.

Simulation

To measure the classification accuracy of a certain exam set, data need to be collected. Item parameters are estimated using an appropriate IRT model. This supposes that one exam measures a certain ability or competency. Subsequently, the covariance matrix for the entire exam set can be estimated. Given this covariance matrix and the item parameters, a latent ability distribution for the population can be built. From either the given distribution or the estimated latent ability distribution, 5,000 true latent abilities are drawn. This enables the researcher to know the true ability and thus the true classification. In the case of an exam set, latent ability vectors are drawn, from which the true classification of the exam set is determined.

Given the latent ability vectors and the item parameters, for each item the probability of a correct answer is computed. Using these probabilities, 5,000 item answer vectors are randomly drawn. From the answer vectors, the observed classification on the exam set is determined. Given the 5,000 observed and true classifications, the classification accuracy is simply determined by the percentage of correctly classified students.

Varied Influences

The ability distribution is determined through the IRT model as estimated. This distribution is supposed to approximate the population ability distribution. In the case of an exam set, the distribution is always a multivariate one. Three variations of the ability distribution were used in this chapter: the estimated empirical distribution, a centered distribution, and a standardized distribution. The empirical distribution was solely estimated from the data sample, described in the next section. Subsequently, the empirical distribution was centralized by subtracting the means of each of the exams; furthermore, the observations were divided by the standard deviation, giving a standard deviation of 1. This leaves the distribution with all means 0. Finally, the distribution was normalized.

Standards are the pre-specified cutoff point for a certain exam. There are two types of standards: norm-referenced and criterion referenced. Though these standards are set differently, they are always a known transformation of each other, given a fixed population. Therefore, this chapter varies norm-referenced standards. Bear in mind that if a norm-referenced criterion requires 60% of the points to be obtained, this is no indication that 60% of the students pass said exam. For each of the other varied influences, the standards are varied from 5% to 95% of the students passing a certain exam. The standards are kept the same for each of the exams in the exam set.

Compensation rules specify how the scores on separate exams should be combined into a single decision on the exam set. These rules prescribe how much compensation is allowed between the different exams. Compensation rules come in three different flavors: conjunctive, complementary, and compensatory (see also Van Rijn et al., 2009). Conjunctive rules allow no compensation; students should have a score higher than the standard on each exam. Complementary rules state that the student should score above the standard for a set number of exams.

Compensatory rules allow complete compensation within an exam set, where the average score of the four exams should exceed a certain pre-specified standard. Conjunctive and complementary compensation rules are often used when there is a minimal level of ability or competence required. Compensatory rules, on the other hand, are often used when it can be justified that deficiencies in one competency are compensated with competency in other areas. There are many compensatory rules possible, depending on the number of exams and the leniency shown (Hambleton, Jaeger, Plake, & Mills, 2000). In the simulation all three types are used, representing increasing leniency. The influence of allowing compensation between exams within an exam set is shown this way.

Example

The data sample was taken from vocational education and training (VET) in the Netherlands. The data used is described next in further detail, as well as the particulars of the simulation. Vocational education and training (VET) in the Netherlands is focused on building and assessing competencies (Gulikers, Bastiaens, & Kirschner, 2004). Therefore, it utilizes performance assessments that find their base in a practical work-like setting. Usually, a few performance assessments are grouped together in an exam set for a certification decision. The exam set may or may not include standardized tests. Although the performance assessment is not the only type of assessment used in certification decisions, it is generally an important one; the performance assessment can constitute as much as 90% of an assessment program.

Data Source

Data were collected from a school for vocational education and training in the Netherlands. The exam set used was from business education and leads to a diploma in Sales Clerk Training (VET). In total, 188 students participated in the study (49% female). This was deemed sufficient, since Martineau (2007) shows that around 200 observations are sufficient to use the computed classification accuracy as a reasonable point estimate of the true classification accuracy. Not every student had completed every exam yet. In VET, it is customary to hold the exam only with students that feel they have a good chance of passing the exam. Therefore, a booklet design where this could be incorporated was used.

The exam set used consists of four performance assessments. Each performance assessment consists of a set of observations. These observations were taken as the items on which the IRT model for the simulation was based. The empirical competence distributions of both data sets were positively skewed and positively kurtosed, compared with the normal distribution (see also Figure 2).

Setup of the Simulation

Since the standards were set differently for each simulation, a way of consistently changing the grading system was devised. First, an estimation of the latent competence level was simulated. Based on that, the grades were divided over percentiles equidistant on both sides of the specified standard. For instance, where the normative standard was on 50% of the students passing, the cutoff percentages were 90, 75, 50, 25 and 10% passing. The students falling in between two given percentiles were given a grade. Some students scored exactly on the percentile. They got the benefit of the doubt and received the higher grade. This conforms to current practice in actual examinations.

The simulated observations were scored with six different grades from 0 to 5. The grade 3 was taken as the passing grade. To examine how the standards set influence the amount of misclassification, a simulation was run for every standard between 5% passing and 95% passing, with 12.5% increments, except for the first and last increment (7.5%). A simulation with these standards was run for every variation of ability distribution and compensation rule.

The data consist of polytomous items, as described above. Therefore, they were analyzed using a polytomous IRT model, the polytomous OPLM model (Verhelst, Glas, & Verstralen, 1993). This model gives the item category response function, describing the response to item i, in which the probability of observing $X_i = j$ as a function of the ability vector, is given by

$$\psi_{ij}(\theta) = Pr(X_i = j|\theta) = \frac{\exp\left[a_i\left(j\,\theta - \sum_{g=1}^{j}\beta_{ig}\right)\right]}{1+\sum_{h=1}^{m_i}\exp\left[a_i\left(h\,\theta - \sum_{g=1}^{h}\beta_{ig}\right)\right]}, \quad (j = 0,\ldots,m_i), \qquad (1)$$

where $\beta_{ig}, g = 1, \ldots, m_i$ are the item response parameters and a_i is the discrimination parameter for each item i.

Using this model, the item parameters per exam were estimated, creating a latent ability distribution under the assumption of a normally distributed ability for each exam. From the four separate ability distributions, a covariance matrix of the latent abilities was estimated. This multivariate normal latent ability distribution is the basis of the simulation. For each of the 5,000 replications, a true ability vector θ_j for replication j, was randomly drawn from the multivariate ability distribution. Given θ_j, the true pass-fail status of each replication was determined.

Subsequently, the response process was imitated by generating a response vector r_j for each replication j. Four randomly drawn numbers from a uniform distribution form a vector u_j, where for each element it holds that $r_j = m \text{ if } u_j \leq P(X_i = m|\theta_j)$. The generated response vectors r_j are used in equation (1) to estimate ability vector $\hat{\theta}_j$, from which again a pass-fail status is gauged. Next, the vectors θ_j and $\hat{\theta}_j$ can be compared to compute the total amount of misclassification. Although the response vectors in the example each consist of four responses, only the pass-fail status of the entire exam set was compared. The pass-fail status was always subject to the compensation rule used in that set of replications.

Four different compensation rules were compared. One conjunctive rule, two complementary rules, and one compensatory decision rule were studied. Although the exam set should always average 3 or higher, there are many restrictions on the separate exams that can be set to influence the diploma decision. Using the conjunctive rule, students will only receive a diploma when they pass all their exams. Thus, every exam should be graded 3 or higher before a student receives a diploma. On the other side of the spectrum is the completely compensatory decision rule, where students obtain their diploma when they have an average on the exam set that is above or equal to the passing grade 3. There are no additional restrictions for a pass on the exam set.

In between are two complementary rules, where students are allowed to compensate one or two deficiencies, respectively. A deficiency is defined as the number of points scored below the specified passing grade. A 2 is one deficiency, a 1 counts as two deficiencies, and a 0 counts

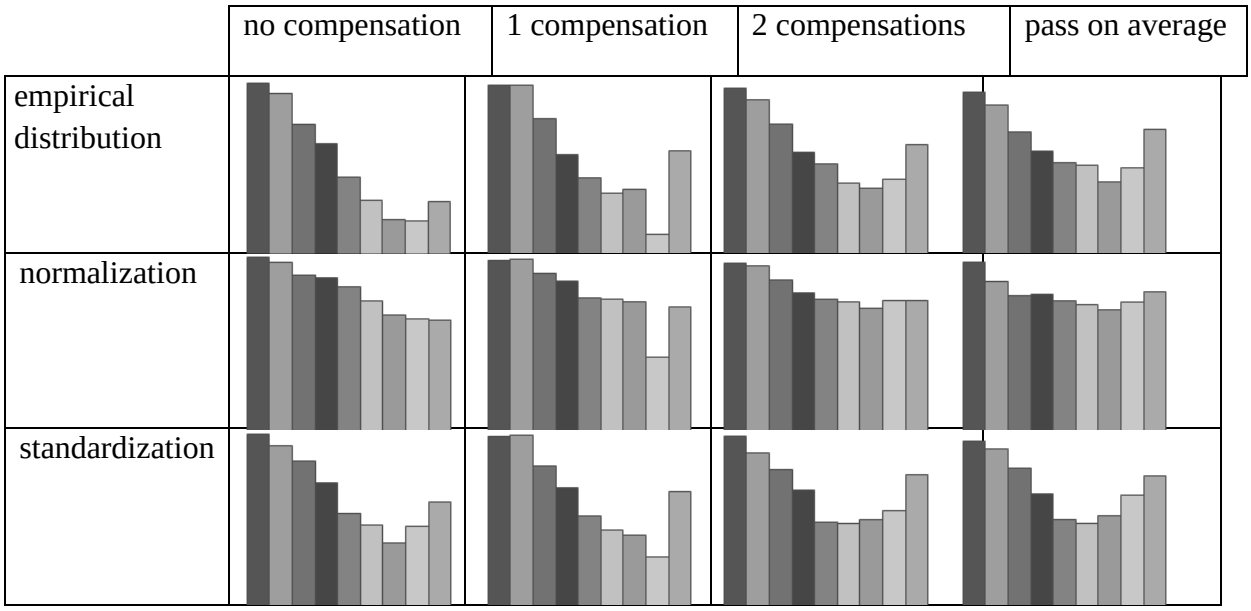
as three deficiencies. With one deficiency allowed in the exam set, a 2 on one of the exams could be compensated with a 4. Compensating a 1 with a 5 is not allowed.

When two compensations are allowed, both a 1 and a 2 can be compensated, given high enough grades on the other exams. In this specific example, the complementary rule that allows three deficiencies is equal to the compensatory rule.

Results

This section reviews the results of the simulations. Knowing what impacts classification accuracy helps to make decisions regarding the examination process. It is important to remember that the classification accuracy reflects the accuracy of the certification decision. First, the influence of the distribution is discussed. Second, the impact of changing normative standards is introduced. Finally, the results regarding different decision rules are discussed.

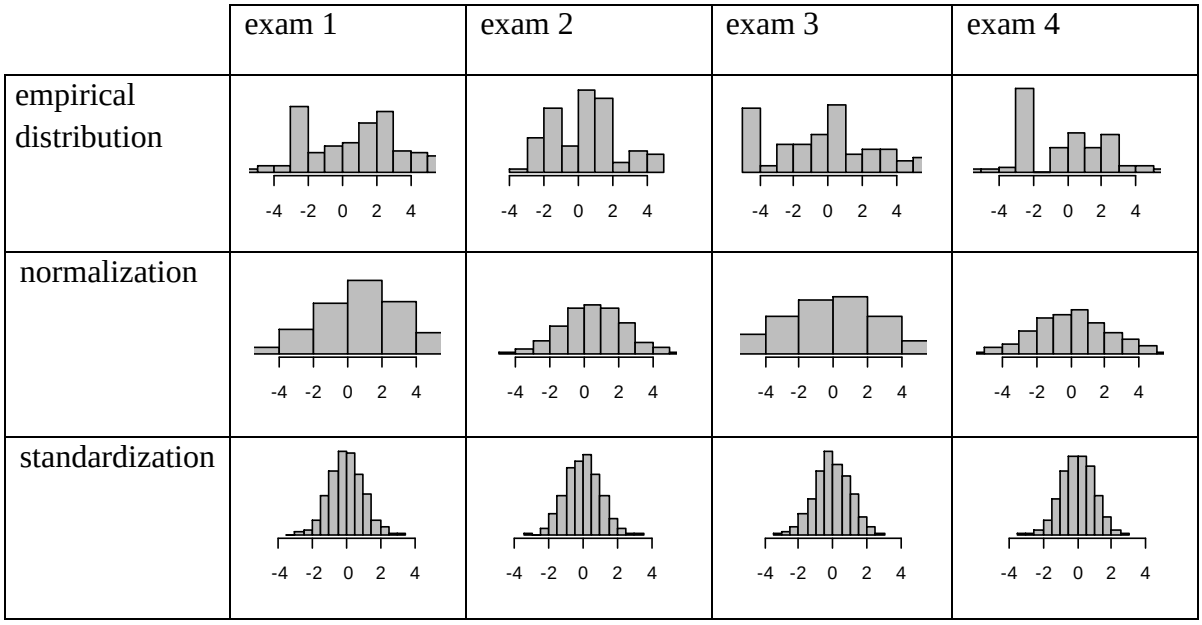
It has been shown that the shape of the ability distribution influences classification accuracy of single exams (Holden & Kelley, 2008). The simulations demonstrate that a similar influence is found for exam sets, as shown in Figure 1. When comparing the three rows of the figure, it becomes apparent that the more the ability distribution of the population follows the assumptions of the model used to compute the classification accuracy, the higher the accuracy becomes. There is a small increase in average classification accuracy from the empirical ability distribution (95%) to the normalized and standardized distributions (both on average 96%). However, visually it is immediately apparent that this is a very consistent result over all the other conditions. This implies both that the model is better in estimating the classification accuracy and that when the model fits the ability distribution of the population, fewer misclassifications are made. These results were expected.



Note: The height of the columns corresponds to the percentages correctly classified students. The scale of the vertical axis is from 88% to 100% accuracy. Each bar plot contains all standards.

Figure 1 Classification accuracy per condition for all standards

A second part of the study looked at the influence of the standards of the exams on classification accuracy of the exam set. In Figure 1, the effect of the standards is visible by inspecting the 12 single graphs. From left to right, each graph shows the standards from 5% obtaining a diploma to 95% obtaining a diploma. Within each graph, the bars show a line with a dip. The dip consistently occurs around the point where the top of the ability distribution of the target population is. This is as expected.



Note: This is a multivariate distribution. The four columns represent the four different exams that make up the exam set used. Rows represent the different distributions. From top to bottom: empirical distribution, normalized distribution, and standardized distribution.

Figure 2 Information on the ability distribution for all distributions

As can be seen in Figure 2, the empirical distribution shows an irregular pattern, resulting in an irregular multivariate distribution. The normalized distribution shows more evenly spread, but still slightly skewed, distributions. This translates into a flatter classification accuracy pattern than for the empirical distribution. The larger dip in the classification accuracy pattern of the standardized distribution coincides with the observed standardized distribution, which is more kurtosed than the normalized distribution.

The highest accuracy is found nearer the extreme standards (where everyone fails every exam or everyone passes every exam) and the lowest accuracy at the top of the ability distribution. Extreme standards, where either nearly all students pass or nearly all fail, are usually not desirable in educational settings. In certain settings, for instance when selecting a top-ten group of students or candidates, extreme standards may apply.

There are two reasons for the coinciding of the dip and the ability distribution. Especially in VET, it is well known what students should know or be able to do for each exam. This undoubtedly leads to “teaching to the test,” a phenomenon where students and teachers work

toward the level the test asks, the standards (Popham, 2001). Some might say that this is a problem. However, when the standards are set at an adequate level, this is not necessarily true. The result should be adequately trained starters and if the standards of the exams are adequate for that goal, it might even be desirable to teach to the test. Should institutes for learning or examination desire a higher classification accuracy, they may consider setting a level of learning higher than the level of examination.

The study also investigates the influence of compensation rules. Figure 1 shows that an exam set in which compensation is allowed leads to higher classification accuracy. When looking from left to right, all classification accuracies show increased classification accuracy, with the increasing freedom of compensation. The third row shows this best. In the last graph, the classification accuracy for each standard is comparable. For standards below 50% of students passing, the compensation rules have a less pronounced effect. This seems due to the ability distribution. The classification accuracy for different compensation rules converges on both sides of the dip. Since a ceiling effect occurs, the classification accuracy does not converge on the higher standards. It must be brought to mind that the compensation is between exams, not within exams. The condition "no compensation" means that each exam must have been finished with a passing grade, not necessarily with a perfect score.

Table 1 Classification accuracy in percentage of correctly classified students

		real standard	average of all standards
empirical distribution	no compensation	88%	94%
	1 compensation	91%	95%
	2 compensations	92%	95%
	pass on average	93%	96%
normalisation	no compensation	97%	98%
	1 compensation	96%	97%
	2 compensations	96%	98%
	pass on average	97%	97%
standardisation	no compensation	96%	96%
	1 compensation	96%	96%
	2 compensations	96%	96%
	pass on average	98%	96%

When just looking at the empirical distribution, set at the currently used standards (see Table 1), the influence of compensation shows that each step of compensation adds some accuracy. No compensation leads to a classification accuracy of 88%.
Obviously, this leaves 12% of the students to be misclassified. Allowing some compensation increases classification accuracy for this exam set (1 compensation: 91%; 2 compensations: 92%).

However, the highest classification accuracy is achieved using full compensation (pass on average: 93%). This implies that nearly half the misclassified students of the no compensation scenario are correctly classified under full compensation. This could be a reason to implement a higher level of compensation. It is interesting also to note that the biggest increase in classification accuracy is found going from no compensation at all to at least some compensation (difference of 91% - 88% = 3%). Depending on the amount of students partaking in the exams, it may be worthwhile to consider at least partial compensation for the exam set. These results had been anticipated as well, since allowing compensation essentially lengthens the exams within the exam set and therefore it is expected that the exam set is better able to classify the students.

Discussion

Most research focuses on the quality of individual exams. However, the important decisions are usually based on an exam set. Therefore, it seems more appropriate to focus on the quality of exam sets. Much has been written about classification accuracy of single exams; there is literature about several ways of computing classification accuracy and there is literature regarding the influences on this classification accuracy.

A few well-known influences are the shape of the ability distribution of the measured ability in the target population, the standard that has been set for the exam, and the allowance of compensation within the exam. It was shown how the classification accuracy of exam sets is influenced by ability distribution, standards, and compensation. This was done using an example from vocational education and training (VET). The classification accuracy for an exam set is computed using item response theory (IRT) simulation. The outcomes indicate that classification accuracy for exam sets is influenced in a way similar to the classification accuracy for single exams. Of course, this stays partially contingent on the quality and number of the exams that together make up the exam set.

Classification accuracy is high when all tests from an exam set have equal and standardized ability distributions. To some extent, schools may exert some influence over the ability distribution. However, researchers or test developers do not have this possibility. Nevertheless, this does imply that concepts that are stable and normally distributed may be tested with higher classification accuracy. Test developers may want to research the concepts they plan to test beforehand, in order to assess whether they can accurately test them.

It is neither always possible nor always desirable to create exam sets that fit the ability distribution of the target population. In some cases, the target population is unknown. Other constraints may be time and money to obtain a distribution. Furthermore, ability distributions may shift over time. Even when the target population and its ability distribution are known, it may change in the time between measuring the distribution and taking the exams. Finally, it is not expected that the ability distribution is completely normal, since most exams in VET are designed with a ceiling effect in mind.

Furthermore, extreme standards (where few or no students pass or fail an exam) increase classification accuracy. In practice, however, it is not common to test far below or above the average ability of the target population. This is especially true in educational settings. Furthermore, developing exam sets that are designed to test outside the ability distribution of the target population do not yield much information. In certain situations, one may still decide to develop such an exam set. For instance, one may want to ensure that all students have a certain basic proficiency, where it is expected that all students easily surpass this proficiency. Measurement is likely to be more accurate if the question of how proficient students are is subordinate to the question of whether all students are proficient enough.

Finally, allowing compensation increases classification accuracy. Compensation, either within a single exam or encompassing the entire exam set, is a property of the exam that researchers or test developers nearly always have an influence on. There are various reasons why one may want to allow or disallow compensation. This discussion limits itself to reasons for allowing and disallowing compensation in exam sets. Reasons for allowing compensation include increasing classification accuracy. Furthermore, when individual exams are short, and they tend to be in VET, it seems reasonable to partially negate the effects of measurement error with the

allowance of compensation between exams. Moreover, intuitively it may feel unfair for a good student to be denied a diploma should she fail just one exam.

Nevertheless, there are arguments in favor of conjunctive exam sets. In some vocations there are at least certain parts of the examination that should be passed in any condition. A nurse that is incapable of inserting an IV needle should not be allowed to practice. In addition, conjunctive measurement may be the fastest, and thus cheapest, way of examination. Besides, it may be the easiest rule to explain to the students, giving them the required insight in their assessment.

On the whole, it is advisable to compute classification accuracy for exam sets. It gives the researcher or test developer insight into the quality of the exam set, rather than just the separate exams. One especially gains insight into how many students are disadvantaged by the method of examination. Moreover, when the quality of the separate exams is difficult to assess, classification accuracy is an elegant measure for gauging the quality of the examination.

On the other hand, computing classification accuracy is rather expensive. It costs time and requires a fair amount of skill on the part of the researcher or test developer, not to mention the lack of easily accessible software. It may only be worth investing the resources when stakes are high, for instance in the case of certification exam sets, or when the quality of the exam set is highly disputed. Of course, when the resources are readily available, classification accuracy sure seems a great measure to add to the findings on exam set quality.

References

Clauser, B. (2000). Recurrent Issues and Recent Advances in Scoring Performance Assessments. *Applied Psychological Measurement, 24*(4), 310-324.

Dochy, F. (2009). The Edumetric Quality of New Modes of Assessment: Some Issues and Prospects. In G. Joughin (Ed.), *Assessment, Learning and Judgement in Higher Education* (pp. 85-114). Springer Science.

Gatti, G. G., & Buckendahl, C. W. (2006). On Correctly Classifying Examinees. In *Annual Meeting of the American Educational Research Association* (San Francisco, CA). Retrieved April 26, 2011 from http://www.unl.edu/buros/biaco/pdf/pres06gatti01.pdf.

Gulikers, J. T. M., Bastiaens, T. J., & Kirschner, P. A. (2004). A Five-Dimensional Framework for Authentic Assessment. *Educational Technology Research and Development, 52*(3), 67-85.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting Performance Standards on Complex Educational Assessments. *Applied Psychological Measurement, 24*(4), 355-366.

Hambleton, R., & Novick, M. (1973). Toward an Integration of Theory and Method for Criterion-Referenced Tests. *Journal of Educational Measurement, 10*(3), 159-170.

Holden, J. E., & Kelley, K. (2008). *Effects of Misclassified Data on Two Methods of Classification Analysis: A Monte Carlo Simulation Study*. Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.

Huynh, H. (1990). Computation and Statistical Inference for Decision Consistency Indexes Based on the Rasch Model. *Journal of Educational Statistics, 15,* 353-368.

Kane, M. (1996). The Precision of Measurements. *Applied Measurement in Education, 9*(4), 355-379.

Lee, W. C. (2008). *Classification Consistency and Accuracy for Complex Assessments Using Item Response Theory*. Iowa City: Center for Advanced Studies in Measurement and Assessment.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, Performance-Based Assessment: Expectations and Validation Criteria. *Educational Researcher, 20*(8), 15-21.

Livingston, S. A., & Lewis, C. (1995). Estimating the Consistency and Accuracy of Classifications Based on Test Scores. *Journal of Educational Measurement, 32,* 179-197.

Martineau, J. A. (2007). An Expansion and Practical Evaluation of Expected Classification Accuracy. *Applied Psychological Measurement, 31*(3), 181-194.

Popham, W. J. (2001). Teaching to the test. *Educational Leadership, 58*(6), 16-20.

Schulz, E. M., Kolen, M. J., & Nicewander, W. A. (1999). A Rationale for Defining Achievement Levels Using IRT-Estimated Domain Scores. *Applied Psychological Measurement, 23,* 347-362.

Swaminathan, H., Hambleton, R. K., & Algina, J. (1974). Reliability of Criterion-Referenced Tests: A Decision-Theoretic Formulation. *Journal of Educational Measurement, 11*,263-268.

Van Rijn, P., Béguin, A., & Verstralen, H. (2009). Zakken of Slagen? De Nauwkeurigheid van Examenuitslagen in het Voortgezet Onderwijs. (Pass or Fail? The Accuracy of Exam Results in Secondary Education) *Pedagogische Studiën, 86,* 185-195.

Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1993). *OPLM: One parameter logistic model.* Computer program and manual. Arnhem: Cito.

Verstralen, H. (2009a). *Quality of Certification Decisions.* Arnhem: Cito.

Verstralen, H. (2009b). *Accuracy of Exams: CTT and IRT Compared.* Arnhem: Cito