ON PRACTICAL ISSUES OF MEASUREMENT DECISION THEORY
An Experimental Study

Jiri Dvorak
Scio, s.r.o., Pobrezni 34, Prague, Czech Republic

Keywords: Educational Measurement, Test Evaluation, Decision Theory, Simulation, Calibration.

Abstract: In Educational Measurement field Item Response Theory is a dominant test evaluation method. A few years ago Lawrence Rudner has introduced an alternative method providing better results than IRT in some cases of measurement. However, his method called Measurement Decision Theory did not get much of interest in the community. In this article we would like to give MDT some of the focus we believe it deserves. In particular we are focusing on the practical issues necessary to successfully implement MDT into a daily life of Educational Measurement. We will summarize classification abilities. After that in the main part of this paper we will explain in depth calibration process which is a crucial part of MDT implementation. A basic calibration process will be described as well as its characteristics. Then, as a main result, an improvement of this basic process will be introduced.

1 INTRODUCTION

The most widely used test evaluation method at this time is the Item Response Theory (IRT). IRT provides great results in estimating the ability level of a tested person. Unfortunately such outcomes are not always applicable. Lots of testing problems are pass/fail problems: HR services, professional certifications, high school or university entrance exams etc. Other tests have to compare person's skills to a given standard defining a set of groups (categories/grades) an examinee could belong to (e.g. CEFRL certification, school grades or state assessments in some countries). These kinds of tests are intended to classify examinees into given (and defined in advance) groups - categories. The purpose of many of today's tests is rather classification than ability estimation. This approach is not new. Even Cronbach and Gleser in their book (Cronbach and Gleser, 1957) argue that the ultimate purpose of testing is to arrive at classification decisions.

Rudner in (Rudner, 2002; Rudner, 2009; Rudner, 2010) discusses main features of IRT usage in solving classification problems. He argues that since classification is a different (and in many ways simpler) task than ability estimation and IRT is fairly complex and relies on a several restrictive assumptions, we should find a more suitable evaluation method intended directly for classification. Rudner then presents educational testing based on the classification named Measurement Decision Theory (MDT). We would like to recall main principles of MDT in the next section (2).

Although MDT has been known for about ten years and even its background was discussed as soon as 1970s (Hambleton and M, 1973; van der Linden and Mellenbergh, 1978), it remains out of the main focus of measurement community. In this paper we would like to give MDT some of the attention we think it deserves. We present a brief overview of efficiency of MDT (section 4) and especially a kind of guideline (an application-ready process) of item parameters estimation in section 5. We recognize lack of research in both topics as one of the main reasons why MDT is used so rarely.

2 BACKGROUND (MEASUREMENT DECISION THEORY)

Measurement Decision Theory (MDT) is a test evaluation method intended to classify examinees. MDT was introduced by Rudner in (Rudner, 2002) and revised in (Rudner, 2009). In his papers Rudner has proven that MDT is simpler and more efficient in classifying examinees than cut-point based IRT classification.

Dvorak J..
ON PRACTICAL ISSUES OF MEASUREMENT DECISION THEORY - An Experimental Study.
DOI: 10.5220/0003901300940099
In *Proceedings of the 4th International Conference on Computer Supported Education* (CSEDU-2012), pages 94-99
ISBN: 978-989-8565-07-5

MDT is nothing more than Naive Bayes Classifier (NBC), a well-known classifier from Artificial Intelligence. Classifiers are algorithms intended to classify objects (examinees in our case of Educational Measurement) according to their attributes (responses to test items) into a pre-defined set of classes/groups/categories. NBC overcomes most of the prerequisites of IRT (especially unidimensionality of tested domain) assuming only a local independence of items similarly to both IRT and CTT.

2.1 Method

2.1.1 Basic Definitions

Def.: Let M be a set of categories and $m_k \in M$ the k-th category. Let U be a set of items and $u_i \in U$ the i-th item. Let Z be a set of examinees and $z_j \in Z$ the j-th examinee.

Def.: Let $P(m_k)$ be a probability of randomly selected examinee belonging to a category $m_k \in M$ and let $\vec{p} = (P(m_1), P(m_2), \ldots, P(m_k))$.

Def.: Let $P(u_i|m_k)$ be a probability of correct response of an examinee of category $m_k \in M$ to the item $u_i \in U$ and let $\vec{p_i} = (P(u_i|m_1), P(u_i|m_2), \ldots, P(u_i|m_k))$ be a vector of parameters of item $u_i \in U$ (i.e. calibration of item $u_i \in U$).

2.1.2 Method Description

Priors. MDT requires us to know in advance sets M and U and two other priors. The first one is a vector of distribution of categories in population $\vec{p}$ and the second one is a set of parameters of items $P = \{\vec{p_i} | u_i \in U\}$.

Observations. Observations obtained from a test for a single examinee are represented by a vector of his/her responses $\vec{z_j} = (z_{j1}, z_{j2,} \ldots, z_{ji})$ where $z_{ji} \in \{0,1\}$ for incorrect/correct response. Let $R = \{\vec{z_j} | z_j \in Z\}$.

Classification. Let's describe the classification process by a function $F : (\vec{p}, P, R) \rightarrow \vec{c}$ where $\vec{c} = (c_1, c_2, \ldots, c_j)$ is vector of categories $c_j \in M$ such as $c_j = m_k \iff$ examinee z_j belongs to category m_k. Function F could be rewritten into a vector of simpler functions $F(\vec{p}, P, R) = (f(\vec{p}, P, \vec{z_1}), f(\vec{p}, P, \vec{z_2}), \ldots, f(\vec{p}, P, \vec{z_j}))$ defined in equation 1 and subsequent equations 2, 3 and 4.

$$f(\vec{p}, P, \vec{z}) = m \in M; P(m|\vec{z}) = \max_{m_k \in M} (P(m_k|\vec{z})) \quad (1)$$

$$P(m_k|\vec{z}) = n_c P(\vec{z}|m_k) P(m_k) \quad (2)$$

$$n_c = \frac{1}{\sum_{m_k \in M} P(\vec{z}|m_k) P(m_k)} \quad (3)$$

$$P(\vec{z}|m_k) = \prod_{\{i|z_i=1\}} P(u_i|m_k) \cdot \prod_{\{i|z_i=0\}} (1 - P(u_i|m_k)) \quad (4)$$

Note that function F is application of Bayes' Theorem. Equation 4 implies the "naive" assumption of local independence of responses to items. n_c used in equation 2 and defined in equation 3 is a normalizing constant ensuring $\sum_{m_k \in M} P(m_k|\vec{z}) = 1$.

3 METHODOLOGY

Our study is based on results of experimental applications of MDT performed in simulated environment. In this section we summarize essential parts of simulation.

Def.: Since simulation often uses randomness we define a random function $RAN : (S, \vec{v}) \rightarrow s \in S$ such as s is selected randomly from S with respect to the probability distribution vector $\vec{v} = (v_1, v_2, \ldots, v_{|S|}), \sum v_i = 1$.

3.1 Test and Item Generator

A test is a set of randomly generated items. Therefore, an item generator is an essential part of a simulation engine. Our model represented by function $GI : \emptyset \rightarrow \vec{p_i}$ is based on two main assumptions. At first it assumes that categories represent sequential grades: for each item u_i and each m_k stands $P(u_i|m_{k-1}) < P(u_i|m_k) < P(u_i|m_{k+1})$. At second that items are quite good: $max(P(u_i|m_{k+1}) - P(u_i|m_k)) \in \langle 0.2, 0.6 \rangle$. Function $GI()$generates $\vec{p_i}$ of random elements $P(u_i|m_k)$ with respect to this assumptions.

4 ACCURACY OF CLASSIFICATION

Instead of recalling Rudner's experiments comparing IRT and MDT we are focusing on practical issues of

MDT. At first we would like to show a relationship between classification accuracy and the number of items or categories respectively. In both cases we are interested in theoretical limits (given quality of items) of accuracy. The classification is performed on actual parameters of items not on the estimated parameters.

Following experiments share common framework. A single experiment is repeated a few hundred times and then statistical characteristics of results of the set of experiments are evaluated. Overall results are presented as a so-called *box-graph* where around a horizontal line representing median is a box showing first and third quartile with whiskers as extreme values. Box-graphs show both the most likely results (medians) and the stability of results (quartiles and extremes).

4.1 Experiments

In this section we present a framework common to experiments following in section 4.2.

Def.: Let's have a given number of categories m, number of items u and number of examinees $z = 200$ defining sets M, U and Z.

Step 1. Let's have a set of parameters of items $P^U = \bigcup_{u_i \in U} GI()$ of actual parameters of items in U, vector of categories $\vec{c}^Z$ such as $c_j^Z = RAN(\{1, 2, \ldots, m\}, \vec{p})$ where $\vec{p} = \left(\frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m}\right)$ (note we are assuming equal distribution of $\vec{p}$), where examinees $z_j \in Z$ belong to and a set R of responses such as $z_{ji} = RAN\left(\{0,1\}, \left\{1 - P^U\left(u_i | c_j^Z\right), P^U\left(u_i | c_j^Z\right)\right\}\right)$.

Step 2. Let $\vec{c} = F\left(\vec{p}, P^U, R\right)$.

Step 3. Let classification error rate $e = E\left(\vec{c}, \vec{c}^Z\right)$ where function $E = (\vec{v}, \vec{w})$ is defined by equation 5.

$$E(\vec{v}, \vec{w}) = \frac{\left|\left\{j | v_j \neq w_j\right\}\right|}{|\vec{w}|} \quad (5)$$

4.2 Results

Here we show results of two sets of experiments. The first one with setting $m = 5$ and $u = (10, 20, 30, 40, 50, 60, 70, 80, 90, 100)$ is shown in Figure 1-left . We can see that the error rate falls with the number of items not only in the sense of the most likely result but also in the sense of stability. The same effect could be seen in Figure 1-right where a similar set of experiments with $m = 8$ is presented.

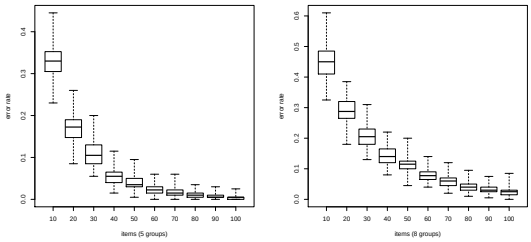

Figure 1: Accuracy of classification vs. number of items (5(left)/8(right) categories, ideal parameters = theoretical limits).

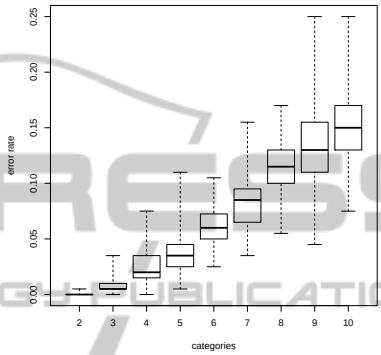

Figure 2: Accuracy of classification vs. number of categories (ideal parameters = theoretical limits).

Figure 2 shows results of an experiment of setting $u = 50$ and $m = (2, 3, 4, 5, 6, 7, 8, 9, 10)$. The accuracy of classification significantly decreases with the increasing number of categories.

In Figure 3 we are presenting an overview of theoretical limits of calibration. We can see how many items we need for a given number of categories to obtain $e < 0.1$. More precisely, the figure shows minimal number of items for a given number of categories when median of error rates was below 0.1.

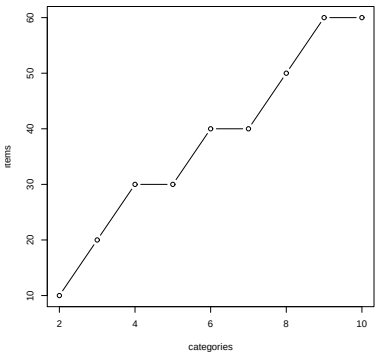

Figure 3: Number of items needed to get median of classification errors below 0.1 for given number of categories.

5 CALIBRATION

In (Rudner, 2002; Rudner, 2009; Rudner, 2010) Rudner spends only a few words talking about MDT calibration (estimation of priors - vector $\vec{p}$ and set P). But for practical purposes calibration process is essential. In this section we are presenting methods of calibration and results of our experiments showing characteristics of calibration process important to implement MDT in real-world testing. In this section we are focused entirely on an estimation of set P because in the worst case, if we were unable to estimate $\vec{p}$, we could set it $\vec{p} = \left(\frac{1}{|M|}, \frac{1}{|M|}, \ldots, \frac{1}{|M|}\right)$ (equally distributed categories in population) without fatal consequences to method precision (see (Rudner, 2009)).

5.1 Basic Calibration

As we have already mentioned MDT is an instance of well-known Naive Bayes Classifier (NBC). NBC is widely used in a scope of Artificial Intelligence where calibration process ("classifier training") is well-developed. In AI there is a "training set" of objects of known attributes as well as their classification. Equivalent to training set in Educational Measurement is pilot testing performed on a set of examinees ("pre-testees") of known classification (typically obtained from external sources e.g. existing certifications). Once we have a set of objects (pre-testees), their attributes (responses to items) and their classifications we are able to compute parameters of attributes (items).

More precisely: Let's have a set of categories M, set of items U , set of examinees Z, set of appropriate responses R and a vector of appropriate classification $\vec{c}$. Our task is to obtain an appropriate set P. Once again we could describe the process as a function B : $(R, \vec{c}) \rightarrow P$. Since P is a set of $P(u_i|m_k)$ elements we could simplify the computation of function B to a computation of each element. In equations 6, 7 and 8 there is a description of evaluation of $P(u_i|m_k)$ in three steps.

$$R_{m_k} = \{\vec{z_j} | \vec{z_j} \in R \wedge c_j = m_k\} \tag{6}$$

$$T_{m_k i} = \{z_{ji} | \vec{z_j} \in R_{m_k} \wedge z_{ji} = 1\} \tag{7}$$

$$P(u_i|m_k) = \frac{|T_{m_k i}|}{|R_{m_k}|} \tag{8}$$

Crucial difference between usage of NBC in Artificial Intelligence and Educational Measurement is the size of a training set. In AI we are typically operating with large training sets even larger than the set of objects we want to classify (see examples in (Caruana and Niculescu-mizil, 2006)). In contrast in Educational Measurement we are very limited in the number of pre-testees. It is an expensive process to recruit persons of known classification especially if we are developing a brand-new test. Therefore there is a strong motivation to keep number of pre-testees as small as possible. Two next sections are dedicated to the analysis of required number of pre-testees (section 5.2) and to the description of a particular calibration improvement technique (section 5.3).

5.2 Items or Categories

Two approaches are possible when describing sufficient number of pre-testees: a per-item (used by Rudner in (Rudner, 2010)) or per-group. In this section we will show which approach is more appropriate.

To answer this question we have constructed two sets of experiments. Experiments are repeated a few times and share a common framework. Results of experiments are again presented as box-graphs.

5.2.1 Framework

Let's have a given number of categories m, number of items u, number of pre-testees z^p, number of examinees $z = 200$ defining sets M, U, Z^p, Z and number of selected items $u^s \leq u$.

Step 1. Let's again have a set of parameters of items $P^U = \bigcup_{u_i \in U} GI()$, vector of categories of pre-testees $z^p_j \in Z^p$ belong to $\vec{c}^{Z^p}$ such as $c_j^{Z^p} = RAN(\{1, 2, \ldots, m\}, \vec{p})$ where $\vec{p} = \left(\frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m}\right)$ (again equal distribution of $\vec{p}$), a set R^p of responses of pre-testees such as $z^p_{ji} = RAN\left(\{0,1\}, \left\{1 - P^U\left(u_i|c_j^{Z^p}\right), P^U\left(u_i|c_j^{Z^p}\right)\right\}\right)$, $U^s \subseteq U$ of u^s_i randomly (equally distributed) selected items $u_i \in U$, and a set of responses of examinees to items of U^s R^s such as $z^s_{ji} = RAN\left(\{0,1\}, \left\{1 - P^U\left(u^s_i|c_j^{Z^p}\right), P^U\left(u^s_i|c_j^{Z^p}\right)\right\}\right)$.

Step 2. Let $P = B\left(R^p, \vec{c}^{Z^p}\right)$ and than let $\vec{c} = F(\vec{p}, P, R^s)$.

Step 3. Let again classification error rate $e = E\left(\vec{c}, \vec{c}^Z\right)$. And let difference of calibration to real parameters $d = D\left(P, P^U\right)$ where function D is defined by equation 9.

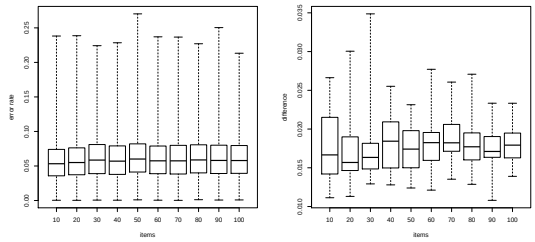

Figure 4: Accuracy of classification(left)/calibration(right) vs. number of items.

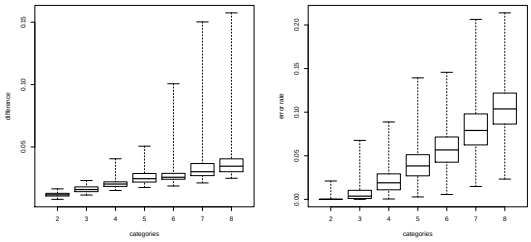

Figure 5: Accuracy of calibration(left)/classification(right) vs. number of categories.

$$D\left(P^1,P^2\right)=\frac{\sum\limits_{u_i\in U,m_k\in M}\left(P^1\left(u_i|m_k\right)-P^2\left(u_i|m_k\right)\right)^2}{m\cdot u} \quad (9)$$

5.2.2 Experiment 1

The first experiment focuses on the per-item approach. This approach suggests that given a fixed number of categories and pre-testees the accuracy of calibration and classification should decrease while the number of items to calibrate increases. We have constructed set of experiments with setting $m=2$, $z^p=20$, $u^s=20$ and $u=(10,20,30,40,50,60,70,80,90,100)$. Note that we are selecting a subset of items from a whole pool to ensure relevant comparison of classification results between experiments with different number of calibrated items with respect to the dependency of classification accuracy to the number of items discussed in section 4. Figure 4-left shows how value of e changes with respect to number of calibrated items. Figure 4-right shows results of d instead of e. As we can see both the accuracy of calibration and the accuracy of classification remain constant.

5.2.3 Experiment 2

The second experiment checks the influence of the number of categories to the accuracy of calibration and classification. The setting of the experiment is now $u=50$, $z^p=30$, $u^s=50$ and $m=(2,3,4,5,6,7,8)$. Figures 5-left and 5-right show the results. Now we can see a very different picture to one seen in the previous figures. The accuracy of calibration as well as the accuracy of classification decreases with the increasing number of groups.

5.2.4 Conclusion

Our experiments have proven that describing number of pre-testees on per-group basis is more appropriate.

5.3 Unknown Objects

Although we have shown that good calibration could be obtained from a relatively small number of pre-testees, the calibration process could be still very expensive and further improvements of the original calibration process are needed. The method we are going to explain was developed experimentally by us independently from the mentioned references.

NBC is used also in document classification. Classification of documents is in many ways similar to testing. In document classification as well as in testing there is a huge amount of objects (documents, examinees) to be classified but it is difficult to obtain a training set. Therefore the training set is typically very small. Nigam et al. (Nigam et al., 2000) took inspiration from (Dempster et al., 1977) and used unclassified objects to improve calibration of NBC. We can use the same approach summarized in the following algorithm:

1. Let's have sets M, U, Z^p, Z, R^p and R and vectors $\vec{p}$ and $\vec{c}^{Z^p}$ (in the notation of previous sections).
2. Let $P_0=B\left(R^p,\vec{c}^{Z^p}\right)$.
3. Let $P_{t+1}=B\left(R,F\left(\vec{p},P_t,R\right)\right)$.
4. Repeat iterative step 3 until terminal condition $P_{t+1}=P_t$ (i.e. $P_{t+1}\left(u_i|m_k\right)=P_t\left(u_i|m_k\right)\ \forall u_i\in U,m_k\in M$) is reached.
5. As a side effect of this calibration classification of examinees we obtain: $\vec{c}=F\left(\vec{p},P_t,R\right)$.

Improvement of both calibration and classification accuracy performed by this iterative calibration algorithm could be seen in Figures 6 and 7 for $m=5$, $u=50$, $z^p=20$ and $z=(100,200,300,400,500)$.

6 CONCLUSIONS

Measurement decision theory is a powerful test evaluation method in cases where we want to classify examinees into a set of pre-defined categories. In this paper we have presented results of a couple of

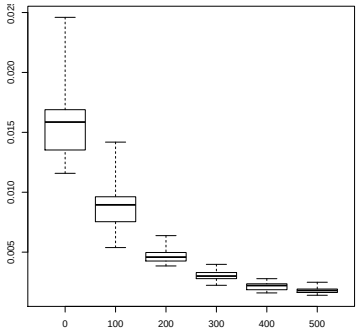

Figure 6: Improvement of calibration accuracy.

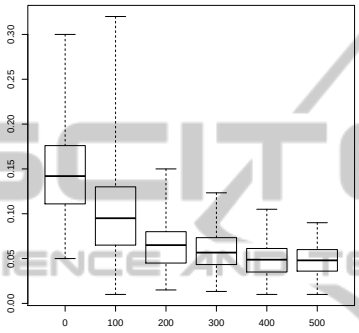

Figure 7: Improvement of classification.

experiments showing some interesting characteristics of MDT. These experiments have followed and expanded on the work (Rudner, 2002; Rudner, 2009; Rudner, 2010) of Rudner. We were focused on practical issues of MDT to give a solid base for future applications of MDT.

We have shown an overview of theoretical limits of classification via MDT in section 4 and a dependency of classification accuracy to number of items and number of categories was discussed. As a main result of this section we have summarized our findings to a direct suggestion of how many items should be chosen to obtain good classification (error rate less than 0.1) results for different number of categories.

In the next section we have focused on the most important obstacle on the path to real-life usage of MDT - the process of calibration of items. We have explained in depth the whole process of simple straightforward calibration of items. Finally we have introduced an improvement to the calibration process which significantly reduces the number of required pre-testees. Experimental results showing the reduction were presented as well.

Based on the results presented in this paper MDT becomes a ready-to-use method.

ACKNOWLEDGEMENTS

This paper was written in the colaboration with my colleagues from Scio (www.scio.cz).

REFERENCES

Caruana, R. and Niculescu-mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *In Proc. 23 rd Intl. Conf. Machine learning (ICML`06)*, pages 161–168.

Cronbach, L. J. and Gleser, G. C. (1957). *Psychological tests and personnel decisions*. Urbana: University of Illinois Press.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. In *Journal of the Royal Statistical Society, Series B, 39 (1)*, pages 1–38.

Hambleton, R. and M, M. N. (1973). Toward an integration of theory and method for criterion-referenced tests. In *Journal of Educational Measurement, 10*, pages 159–170.

Nigam, K., Mccallum, A. K., Thrun, S., and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using em. In *Machine Learning, 39 (2/3)*, pages 103–134.

Rudner, L. M. (2002). An examination of decision-theory adaptive testing procedures. In *Paper presented at the annual meeting of the American Educational Research Association, April 2002*.

Rudner, L. M. (2009). Scoring and classifying examinees using measurement decision theory. In *Practical Assessment Research & Evaluation, 14(8)*.

Rudner, L. M. (2010). Measurement decision theory (a measurement decision theory tutorial). In *http://echo.edres.org:8080/mdt/*.

van der Linden, W. J. and Mellenbergh, G. J. (1978). Coefficients for tests from a decision-theoretic point of view. In *Applied Psychological Measurement, 2*, pages 119–134.