European Journal of Science and Mathematics Education
Vol. 4, No. 4, 2016, 477-489

A comparative analysis of common mistakes in achievement tests prepared by school teachers and corporate trainers

Ali Simsek
Department of Communication Design and Management, Anadolu University, Eskisehir, Turkey
For correspondence: asimsek@anadolu.edu.tr

Abstract
Test development has been an important part of measurement and evaluation in any educational setting, whether its purpose is instruction or training. Both teachers and trainers are expected to have certain level of mastery in developing reliable and valid tests for assessing performance of learners adequately. However, it has often been reported that teachers and trainers are not competent enough to develop such tests. Most teachers have taken at least one university course on educational measurement and evaluation during their undergraduate education. Similarly, a considerable number of corporate trainers attend at least one in-service training seminar on measurement and evaluation as a part of their professional training program. However, teachers and trainers still make serious mistakes in preparing tests for assessing the level of learning and performance properly. This study compares their mistakes based on basic principles of test development. More specifically, the researcher wanted to assess whether the mistakes of teachers and trainers are similar or different. Toward this purpose, a total of 120 instructors (62 teachers and 58 trainers) were selected as participants of the study. A total of 6450 test items in various fields of learning were analysed to make comparisons. Results generally suggest that school teachers and corporate trainers make similar mistakes, although their level of knowledge and skills in measurement and evaluation are different due to prior learning. There is also no difference regarding the subject matter area and the level of education for which the tests were developed. Approximately 60% of items need revisions, half of which appear to have severe mistakes. The most common problems are related to easy test items, route learning, implausible distractors, use of negative questions, hidden cues, illogical order of alternatives, length of sentences for correct answers, disturbing sequence of items, single-mode (text-based) questions, and subjective items. These results, along with some others, have serious implications for programs on teacher education and training of trainers.

Keywords: achievement tests, teacher-made tests, corporate training tests, performance assessment, mistakes in test development, educational measurement and evaluation.

Introduction

Teaching, in essence, is about helping students achieve specified learning objectives. Usually education ministries or school systems identify what is expected of students when they complete a course. The teacher is a facilitator guiding and supporting students accomplish the intended outcomes. The process of educational evaluation is concerned how well students have attained the learning objectives. At least the formative part of this process is undertaken by teachers because the assessment results inform teachers about successes and failures of their students. For this reason, teachers are trained on how to develop and implement appropriate assessment tools as well as how to analyse the results and communicate them with students to improve instruction.

The situation is not very different in corporate training establishments, although there are certain nuances compared to school settings. Trainers are not teaching professionals as school teachers. Instead, they are selected from among the employees and trained through relatively short seminars, which is often called “training of trainers” programs. Within such programs, certain amount of time is spent on performance assessment tools and techniques. This process helps trainers develop a general understanding about the relationship between instruction and evaluation.

Test development has been a crucial part of measurement and evaluation in any educational setting, whether its purpose is teaching or training. Both teachers and trainers are expected to have certain level of mastery in developing reliable and valid tests for assessing performance of learners adequately. Considering that many decisions about learners are made based on their performance on various tests, developing good achievement tests becomes extremely important. For example, students in schools may or may not graduate based on test results. Similarly, employees in corporations may or may not get promotion according to their scores on relevant tests. This makes test development a critical job for teachers and trainers.

Teachers collect and use multiple sources of information to assess learning, they assist students in assessing their own learning, and they communicate assessment results with students, parents, other educators, and the relevant authorities (AFT, NCM, & NEA, 1990). Similarly, trainers decide how to judge the trainee's ability to perform various tasks. It is customary to test the trainee's skills at the end of each unit and also at the end of the training program (FAO, 2014).

Because teachers and trainers are primarily responsible for assessing student learning and evaluating instruction, there is a widespread concern about the quality of assessment. The consensus has been that teachers use a variety of assessment tools or techniques, even though they may be inadequately trained in certain areas of educational assessment (Zhang & Burry-Stock, 2003).

It has often been reported that teachers and trainers are not competent enough to develop achievement tests. As a consequence of this, many teachers and trainers inaccurately evaluate performance of their students. They place their faith in poorly designed tests but the test results may not reflect actual achievement of the specified objectives (Stiggins, 1988). For example, if a teacher-made test consists of excessive number of easy items, the learners will be highly successful, when in fact specified objectives of the curriculum require higher-order competencies in order to be considered successful. The reverse is true when the teacher-made test is extremely challenging and the learners fail, although they have achieved objectives of the curriculum.

Measurement is generally defined as the process by which facts or data are gathered through empirical observations. Teachers and trainers usually measure student achievement by reading what students have written, watching their performances, listening to what they say, and using the basic senses to gather information relevant to the specified objectives. As a closely-related concept, evaluation is a value judgment and it is influenced by the results of measurement (Cangelosi, 1990; Mehrens & Lehmann, 1984).

Teachers in schools and trainers in corporations often use tests as a tool of assessment. The test is a planned measurement by which appropriate opportunities are provided for students to display their performance relevant to specified objectives. Tests are composed of items. Each item on the test first confronts students with a particular task and then provides a means for observing their responses to the task.

There are many kinds of tests, and the rules for developing tests differ according to their kinds. Achievement tests are often developed by teachers. Similarly, performance tests are usually designed by trainers in the corporate world; only the terms are different, substance stays the same. Stages and steps for developing achievement/performance tests are relatively straightforward. They can be indicated as: (1) Specifying learning objectives; (2) Developing a test blueprint; (3) Creating item pools; (4) Synthesizing the test; (5) Administering the test; (6) Scoring the test; (7) Declaring scores; (8)

Reviewing the test results with students. Of course, different tasks are undertaken during each stage and some of the steps involve heavy and complicated work.

Important rules and principles for preparing test items are generally agreed upon in the literature. They vary according item types. Because the most commonly used item type is multiple-choice, guiding principles are listed below (Simsek, 2013; Wiggins, 1992):

- Consider learning objectives
- Provide clear directions
- Judge parallel alternatives carefully
- Control item difficulty
- Avoid overlapping items and alternatives
- Pay attention to correct grammar
- Use plausible distractors
- Avoid ambiguity
- Locate correct response randomly
- Use appropriate number of alternatives
- Avoid unnecessary cues or hints
- Apply correction for guessing
- Use words efficiently
- Avoid subjective or judgmental questions
- Organize alternatives logically
- Avoid "none" or "all" options
- Emphasize negative statements in questions
- Avoid double-barrel items
- Provide the right amount of information
- Avoid implying the correct response

Teachers and trainers can use performance tests to assess complex cognitive processes as well as attitudes and social skills in academic areas. When doing so, they establish situations that allow them directly to observe and to rate learners as they analyse, solve problems, experiment, make decisions, measure, cooperate with others, present orally, or produce a product. These situations simulate real world activities, as might be expected in a job, in the community, or in various forms of advanced training such as in the military, a technical institute, on the job training or college. Performance tests also allow teachers and trainers to observe achievements, habits of mind, ways of working, and behaviours of value in the real world that conventional tests may miss (Borich, 1996, p.634).

Needless to say that basic competencies related to educational measurement and evaluation (particularly test development skills) make of a vital dimension of teachers' or trainers' professional capabilities. Therefore, regardless of their educational settings and subject matter areas, all teachers and trainers are expected to develop reliable, usable, and valid achievement tests complying with fundamental rules and principles, especially after attending relevant and sufficient training programs.

Research on teachers' assessment practices revealed that they were not well prepared to meet the demand of educational assessment due to inadequate training (Hills, 1991; O'Sullivan & Chalnick, 1991). Problems were particularly prominent in performance assessment, interpretation of standardized test results, and grading procedures (Zhang & Burry-Stock, 2003). When using performance measures, many teachers did not define exact levels of performance or plan scoring procedures before instruction, nor did they record scoring results during assessment (Stiggins & Conklin, 1992). Teachers also had trouble communicating test results (Plake, 1993). Many teachers incorporated non-achievement factors such as effort, attitude, and motivation into grades (Griswold, 1993; Hills, 1991) and they often did not apply weights in grading to reflect the differential importance of various assessment components (Stiggins, Frisbie, & Griswold, 1989). Despite the aforementioned problems, most teachers believed that they had adequate knowledge of testing (Kennedy, 1993) and attributed that knowledge to personal experience and university coursework (Wise, Lukin, & Roos, 1991).

Teachers' concern about the quality of classroom assessment varies with grade levels and slightly with subject areas (Stiggins & Conklin, 1992). There is an increased concern among teachers about the improvement of teacher-made objective tests at higher-grade levels; mathematics and science teachers were more concerned about the quality of the tests they produced than were writing teachers (Zhang & Burry-Stock, 2003).

It is reasonable to assume that training instructors on educational measurement and evaluation makes a big difference in developing tests. Zhang and Burry-Stock (2003) reported that regardless of teaching experience, teachers with measurement training report a higher level of self-perceived assessment skills in using performance measures; in standardized testing, test revision, and instructional improvement; as well as communicating assessment results than those without measurement training.

The present study examines whether school teachers and corporate trainers prepare acceptable test items for assessing student achievement. This study extends the current literature by comparing test items prepared by teachers and trainers based on fundamental principles of test development. Such a comparison is meaningful from the point of professionalism in teaching since the nature and the level of measurement training is different for school teachers and corporate trainers. It is expected that the results of the study will have some implications for teacher education programs and the programs for training of trainers.

This study seeks answers to the following questions by comparing common mistakes made in achievement tests prepared by teachers and trainers:

- Do achievement tests developed by school teachers and corporate trainers comply with the required principles of educational measurement?
- Is there a difference between teachers and trainers in applying these principles successfully in developing achievement tests?
- What kinds of common mistakes that teachers and trainers make in preparing test items for measuring student performance?
- Are the mistakes observed in test items severe and require serious improvements before implementation or are they simple and can be neglected?

Methods

Sample

A total of 6450 test items were analysed toward the purpose of this study. These items were prepared by 120 instructors, of this number 62 were teachers in a private school system and 58 were corporate trainers in three different establishments in Western Turkey. The school system had both an elementary school (grades 4-8) and a secondary school (grades 9-11). Gender distribution of the teachers was almost equal with an average teaching experience of 10 years. All the teachers had university degrees. They had taken at least one "measurement and evaluation" course during their undergraduate education.

The first corporate establishment has been functioning in the area of logistics management, the second in the area of automotive manufacturing, and the third in the area of textile design. The first two corporate establishments had corporate academies and the last one had a testing centre for occupational certification. It means that test development was a vital part of these establishments.

Trainers in all the three corporations were selected and trained carefully. They all graduated from a four-year college and came from different backgrounds, mostly engineering, design, and management fields. Approximately two-thirds of the trainers were male, and the average training experience for the whole group was about 5 years. Only a few of them had taken educational assessment course during their university education. Regardless of their academic backgrounds, they were all trained in educational measurement and evaluation as a part of their in-house training seminars.

Procedures

The researcher has served as an educational consultant for all these schools and corporations for reasonable time periods (18 months in each). Part of his responsibilities as a consultant included training teachers and trainers in test development, helping them write test items for assessment of achievement, evaluate their work critically, and provide feedback for possible improvement.

School teachers attended an 80 hours training program, of which approximately 20 hours consisted of sub-training in educational measurement and evaluation. This was a part of the project on accommodating individual differences in education. Sub-training focused on test planning, writing test items, analysing items, calculating reliability, and improving test items in light of student responses. During the sub-training, teachers attended face-to-face instruction, wrote pilot test items, received feedback about them, and corrected items which appeared to be problematic. Then, they wrote actual test items for an achievement test to be used in their regular classes.

Corporate trainers attended "training of trainers" programs which lasted for duration of 64 hours. Each of these programs was conducted separately at different times. Approximately 18 hours of training of trainers programs consisted of sessions on educational measurement and evaluation. During the training, they attended face-to-face instruction, exercised writing test items, received feedback regarding important mistakes, and benefitted from suggestions on possible ways of improvement their items. Then, they wrote actual test items to be used for assessing achievement at the end of instructional modules. These tests were implemented as a part of pilot-testing of the corporate training modules.

Data Analysis

As mentioned earlier, all the 6450 test items were analysed by a panel of three instructional designers who were also sufficiently knowledgeable on educational assessment, one being the researcher. First, the reviewers worked on about certain examples (about 200 items) to check whether they could evaluate the same items consistently. This produced a high agreement among them. Then, they split the items into three for independent evaluations. However, the researcher acted as the principal reviewer so that he reviewed all the 6450 items before writing the meta-evaluation report and briefing the test writers about their mistakes. The overall inter-rater reliability was about .95 among the three reviewers, which means that their reviews were consistent and highly reliable.

Findings

All the test items prepared by school teachers and corporate trainers were analysed from certain points guided by the principles of educational assessment. Results of these analyses were presented in tables and their corresponding interpretations were provided after each table. Table 1 demonstrates the number of teachers and trainers participating in the study and prepared the test items.

Table 1. The Number of Teachers and Trainers Participating in the Study

Status	**n**	**%**
School teachers	62	52
Corporate trainers	58	48
Total	**120**	**100**

The distributions of teachers and trainers in the sample were close to each other (about 50% each), although the number of teachers was relatively higher (n=62 versus n=58). Teachers were distributed almost evenly between elementary (n=32) and secondary (n=30) education. Trainers were from three different areas, namely logistics management (n=30), automotive manufacturing (n=18), and textile design (n=10). The number of trainers in logistics management was close (or slightly higher) to the combined number of trainers in the areas of automotive manufacturing and textile design.

Distribution of the test items according to the types of education (school learning versus corporate training) was presented in Table 2.

Table 2. Distribution of Test Items According to Types of Education

Types of Education	**f**	**%**
School learning	***4818***	***75***
Elementary education	2380	37
Secondary education	2438	38
Corporate training	***1632***	***25***
Logistics management	960	15
Automotive manufacturing	480	7
Textile design	192	3
Grand total	**6450**	**100**

Approximately 75% of the test items were related to areas of school learning; of this 37% was for elementary education and 38% was for secondary education. Approximately 25% of the test items were about corporate training; of this 15% was for logistics management, 7% were for automotive manufacturing, and 3% were textile design. We can claim that these percentages did not jeopardize the comparative analysis because the number of test items prepared by corporate trainers was still high and sufficient for meaningful comparisons (f=1632 versus f=4818).

Distribution of the test items according to grade levels for school learning and certification levels for corporate training was shown in Table 3.

Table 3. Distribution of Test Items According to Grade Levels

Grade Levels	**f**	**%**
Elementary Education	**2380**	**37**
Grade 4	440	7
Grade 5	400	6
Grade 6	500	8
Grade 7	500	8
Grade 8	540	8
Secondary Education	***2438***	***38***
Grade 9	748	12
Grade 10	850	13
Grade 11	840	13
Adult Learning	***1632***	***25***
Basic	1220	19

Development	412	6
Grand total	**6450**	**100**

Within the level of elementary education, percentages of the test items by grades were very close, ranging from 6% to 8%. Within the secondary education, the situation was similar, ranging from 12% to 13%. For corporate training, however, the situation was different. Because most of the training modules were developed for basic vocational training for new employees, the percentage of the test items for this level was almost three times higher than the professional development level.

Distribution of the test items according to the domains of learning (based on Bloom's taxonomy) was provided in Table 4.

Table 4. Distribution of Questions According to Domains

Domains of learning	**f**	**%**
Cognitive	6130	95
Affective	168	3
Psychomotor	152	2
Total	**6450**	**100**

Although the courses and modules for which the test items were developed appeared to have some affective and psychomotor goals, 95% of the test items were prepared for categories of the cognitive domain.

Within the cognitive domain 2097 (33%) items were related to knowledge, 1894 (29%) items were related to comprehension, 1035 (16%) items were related to application, 512 (8%) items were related to analysis, 541 (8%) items were related to synthesis, and 55 (1%) items were related to evaluation.

Affective domain consisted of 168 (3%) of the total number of items. These items were distributed closely among the categories of receiving, responding, and valuing except a few items in the category of organization.

Psychomotor domain consisted of only 152 (2%) of items. These items were mostly related to perception, set, and guided response; with a few items in each of the remaining categories.

As far as the distribution of items among subject domains is concerned, no differences were observed between school learning and corporate training as well as between elementary education and secondary education. However, almost all of the items in the psychomotor domain have come from corporate training modules; the only exception in school learning was Physical Education course. Affective items were mostly related to Turkish History course in school learning and Quality Indicators, Health and Safety, and Design modules in corporate training. There were several courses (i.e. Culture of Religion) for which almost all the items were in the category of knowledge alone.

Considering that both teachers and trainers concentrated heavily on writing test items in cognitive domain, one can assume that the test items generally reflect a cognitive orientation to teaching. Instructors might have thought that achievement tests were appropriate only for assessing cognitive performance. It is also important to note that a great majority of the items were related to learning tasks in the earlier categories of each domain.

Distribution of the test items according to the subject matter areas of school learning was presented in Table 5.

Table 5. Distribution of Questions According to Areas of School Learning

Areas of Learning	**Number of Items**
English	600
Mathematics	400
Culture of Religion	350
Literature	300
Arts	280
Music	250
History	250
German	250
Computer	250
Science	250
Linguistics	200
Geography	200
Social Studies	200
Turkish Language	200
Chemistry	198
Physics	150
Biology	150
Calculus	100
Geometry	100
Physical Education	55
Philosophy	50
Logic	15
Psychology	10
Sociology	10
Total	**4818**

The test items related to school learning cover almost all subject matter areas taught in typical elementary and secondary schools. It appears that the number of items in the area of English (f=600) ranks the top among all. This is because the private school system in the current study had an English preparatory year where English was taught as a second language. The second area that consisted of the largest number of test items is mathematics (f=400). If we add the items for geometry and calculus to this, mathematics shares the top rank with English. The Culture of Religion is the third largest area that consisted of 350 items, followed by Literature with 300 items. In contrast, the areas of Sociology, Psychology, Logic, and Philosophy had the lowest number of test items.

These figures are due to two major reasons: In some cases they show the weights of subject matter areas or courses in the overall curriculum. The situation of English, Mathematics, and Literature reflects this. In some other cases the number of items for an area shows the efforts or disinterests of teachers in assessment. For example, Physical Education teachers prepared a generic set of items for all grade levels, whereas the teachers for Culture of Religion prepared different sets of test items for each grade level. Nevertheless, most of the teachers prepared adequate number of test items consistent with the weights and objectives of their courses as indicated in the curriculum. In any way, English and Mathematics had the highest number of test items, while Sociology and Psychology had the lowest number of items.

Distribution of the test items according to the areas and modules of corporate training was shown in Table 6.

As seen in the table, the test items for corporate training split into three major clusters such as logistics management, automotive manufacturing, and textile design. The logistics management cluster has the highest number of modules (f=24) as well as the highest number of test items (f=960). On the other hand, textile design cluster has both the lowest number of modules (f=7) and the lowest number of test items (f=192). Automotive manufacturing cluster is in the middle with 12 modules and 480 test items.

Since the structures of the training modules are predetermined carefully by the instructional design teams, the numbers of test items in each module ended up equal. Each of the modules related to logistics management and automotive manufacturing had 40 test items. The textile design subcategory showed a small deviation because it had the highest number of 30 items and the lowest number of 25 items.

Table 6. Distribution of Questions According to Areas of Corporate Training

Areas of Learning	**Number of Items**
Logistics Management	***960***
Intro. to Agency Services	40
Marketing and Sales	40
Documentation	40
Customs	40
Waybill	40
Container Management and Control	40
Ship	40
Operations	40
Forwarding	40
Claim Management	40
Ship Operations	40
External Accounts	40
Health and Safety	40
Logistics Law	40
Risk Management	40
Intro. to Port Services	40
Port Services	40
Operations	40
Commercial Tariffs	40
Gate Operations	40
CFS Operations	40
Zone Operations	40
Warehouse Operations	40
English for Agency Services	40
Automotive Manufacturing	***480***
Process Control	40
Job Security	40
ISO 9001	40
Quality Indicators	40
Measurement Design	40
Problem Solving	40
Metal Processing	40
Virtual Analysis	40
Product Development	40

Vehicle Geometry	40
FME Analysis	40
Product Management	40
Textile Design	***192***
Job safety	30
Communication	30
Labour Law	30
Design	27
Marketing	25
Production Techniques	25
Management	25
Total	**1632**

Common mistakes observed in the test items prepared by teachers and trainers are listed comparatively in Table 7 from the most common to the least common problems. Table 7 does not necessarily reflect the most serious mistakes; instead, it demonstrates the most frequently occurred mistakes. Some of these mistakes are severe but others are simple. Severe mistakes are against measurement principles and have a potential to affect test results negatively, while simple mistakes may not comply with measurement principles but do not also have negative effects on test results.

Table 7. Common Mistakes in Test Items

Mistakes	**Teacher**	**Trainer**	**Total f**
Excessive number of items in cognitive domain	4577 (%95)	1553 (%95)	6130 (%95)
Easy test items that do not require thinking	3587 (%74)	1250 (%77)	4837 (%75)
Implausible or illogical distractors	2450 (%51)	780 (%48)	3230 (%50)
Negative questions without attention focusing	1656 (%34)	600 (%37)	2256 (%35)
Hidden cues or hints in statements of questions	1465 (%30)	469 (%29)	1934 (%30)
Illogical/improper listing of alternatives	1350 (%28)	457 (%28)	1807 (%28)
Presenting the answer in the longest alternative	1280 (%27)	460 (%28)	1740 (%27)
Using the same letter option consecutively for the answer	1162 (%24)	451 (%28)	1613 (%25)
Asking subjective and/or judgmental questions	993 (%21)	425 (%26)	1418 (%22)
Not using visuals in questions when appropriate	964 (%20)	326 (%20)	1290 (%20)

Note: Some mistakes are repeated across items so that the totals do not add up.

The highest number of mistakes in test items was related to *domains of learning*. Both teachers and trainers wrote 95% of the items for cognitive tasks, leaving only the remaining 5% for affective and psychomotor domains. This may be understandable for certain courses such as Mathematics or Labour Law. However, the instructors wrote excessive number of cognitive items even for Music or Metal Processing. In general, heavy concentration on cognitive tasks in assessment may be considered a problem when the variety of the courses and modules are taken into account.

It was identified that the three-fourth of the *test items was easy* and related to the lower categories of cognitive domain such as knowledge, comprehension, and application. In fact, the most items were recall and recognition questions, leaving little space for information processing and knowledge construction. This may be in part that the teachers and trainers participating in the current study were employees of private establishments so they wanted to demonstrate that their students were successful, implying that they were good instructors. Pilot tests have shown that the average item difficulty was around .80 which means that four out of five students were able to answer the questions.

Approximately half of the items had *implausible/illogical distractors*. Apparently irrelevant or easily avoidable option in a test item helps respondents skip them immediately and identify the correct response in a short time. Such a process avoids information processing and alternative thinking which eventually help students answer the questions when in fact they do not know what the correct response is. This also reduces item difficulty, which is particularly undesirable when mastery on learning tasks is important.

Teachers and trainers asked so many *negative questions* and used the negative statements in items without attention focusing. Approximately one-third of the items included negative statements and many of them did not highlight the negative parts of the questions. Moreover, a number of negative items were often sequenced consecutively. It should be noted that 3-4 negative items in a row were quite common; some tests even started with negative items against the well-known convention of educational assessment.

Approximately one-third of the items included *hints or cues* that helped students answer some of the questions correctly when in fact they did not know the answer. This problem occurs when the question statement of an item covers helpful indications for identifying the correct response for another item. It is usually suggested that this should be checked across all items and avoided throughout the test. However, it seems that the teachers and trainers writing test items were not careful enough regarding this issue. As a result, students with high test-taking skills obtained an advantage when they were able to detect hints or cues among items.

For many items, *order of options* was not logical or proper. This problem was observed in about 28% of the items. It was particularly apparent in areas such as mathematics and history. In mathematics, for example, it is suggested that the options of an item should be ordered from the smallest value to the biggest value. Similarly, in history, alternatives of an item should be sequenced chronologically. Unfortunately, this was not the case in a considerable number of items and it seemed that teachers and trainers ordered the options as they have come to their minds. One should not forget that random order of options may sometimes provide unnecessary hints for students.

Many test items embedded the correct response into *the longest sentence* so that students were able to identify the answer with little effort. The long sentences normally give the impression that they are more complete and meaningful than the statements in other options so that students review these sentences first to shorten the responding time. The length of a sentence in an option should not be visibly different from others and a careful balance should be established.

One-fourth of items used the *same letter option* consecutively to represent the correct response. It is customary to use the same letter 2-3 times in consecutive order to represent the answer. For example, the correct response may be represented by the letter B for two questions following each other in a test. However, when the answer is represented by letter B for four consecutive questions, this may force students have a second thought. They usually think that they have made a mistake somewhere because four B's in a row are not acceptable. Then, they may even change their response from the correct one to the incorrect one due to this assumption. At the end, reliability of the test is reduced.

Approximately 22 % of the items asked *subjective and judgmental questions* to students. Trainers tended to do this relatively more than teachers. It may be that their students were working adults who had some experiences in their respective fields. Judgmental questions may cause a number of problems in achievement tests, particularly when acceptable indicators of performance are predetermined and straightforward. The items starting with statements like "in your opinion" or "to your best knowledge" may produce personally defendable but academically unacceptable answers. Because the item asks individual view of the respondent, any response should be accepted as correct. Such practices may be appropriate for attitude scales or interest inventories but not achievement tests which measure competence or gains in learning.

Another common problem was that many items did not use *visual stimuli* when feasible. It is understandable that not all items can have visual or auditory stimuli. However, considering that the spectrum of the courses or modules is fairly large, particularly visual stimuli could be used in a number of items. For example, the pictures of instruments could be used in music test; certain movements in sports could be visualized in Physical Education test; types of containers could be shown in Container Management test, or parts of a car could be pictured in Vehicle Geometry test. Unfortunately, test items took the advantage of visualization much less than one expects in this study. As a result, these tests turned out to be full of text-based items, which might have created disadvantages for visual learners.

Results and Conclusions

Generally speaking, approximately 60% of the test items had mistakes that needed to be corrected or improved before administration. While some of them were critical mistakes that required considerable revisions by the item writers or subject matter experts, others could be improved by the experts of educational measurement and evaluation. Of the total number, approximately half of the test items had severe problems that tests could not be administered without excluding these items or correcting them radically. This suggests that most teachers and trainers are still not capable of developing good achievement tests, which is consistent with the findings in the literature. The main reason appears to be inadequate training (Hills, 1991; O'Sullivan & Chalnick, 1991; Zhang & Burry-Stock, 2003).

Common mistakes in test items showed a wide variety ranging from content mistakes (related to subject matter area) to format mistakes (related to measurement). Content mistakes were much less than format mistakes. The most prominent mistake was asking easy items in the cognitive domain. In other words, teachers and trainers asked questions that usually required route learning. They used predominantly text-based questions and negative statements without focusing students' attention. This might have created an advantage for verbal learners and disadvantage for visual learners. They also provided hidden clues that helped those with high test-taking skills, presented implausible options which were not even worth to think, placed the correct response into the longest sentence implying the answer, and used disturbing order of alternatives as well as representations of correct responses instigating students have a second thought about their response. Although the items were prepared for multiple-choice tests, the teachers and trainers also asked considerable number of subjective and judgmental questions. Along with many other minor mistakes, it can be concluded that the test items analysed in the present study have not complied with the basic principles of educational measurement and evaluation. This comes as no surprise considering similar results in the literature (Griswold, 1993; Kennedy, 1993; Stiggins, 1988).

It appears that the mistakes that teachers and trainers made in preparing their test items did not differ significantly. The largest difference was within the limit of 5 percentile points for asking subjective items. It seems that corporate trainers ask more subjective or judgmental questions compared to school teachers. Normally, one expects that school teachers are better test developers because they receive more education in the field of educational measurement and it is also a vital part of their professional activities. However, this study shows that they are not skilful enough in writing test items that measure student learning both in scope and depth. Similarly, corporate trainers are not better test developers either. Taking together, these results suggest that both teachers and trainers lack adequate training, although they believe that they have adequate knowledge of testing

(Kennedy, 1993) which can be attributed to personal experience and university coursework (Wise, Lukin, & Roos, 1991).

There are no visible differences in mistakes according to levels of schooling and areas of teaching. One expects that, due to prior training and professional expertise of school teachers compared to corporate trainers, test items related to areas of school learning would have less mistakes. However, this is not the case in this study. In fact, corporate trainers of certain areas prepared better test items. It appears that both the numbers and the types of mistakes are more related to competence of individual teachers or trainers rather than their subject matter areas. For example, a teacher in the field of literature or history wrote better test items than the teachers of mathematics or physics, although the latter areas are closer or more relevant to the field of measurement and evaluation. This runs contrary to the findings of some studies (Stiggins & Conklin, 1992; Zhang & Burry-Stock, 2003). It may be due to the fact that some tests were prepared in collaboration of teachers in the same area of teaching. For example, English tests were prepared together by all English teachers in the school, at least they had a chance to review and approve each other's work. Similarly, corporate trainers reviewed other modules and exchanged ideas with their colleagues.

Further research is needed in this area. First, tests developed by well-trained instructors in the field of educational measurement and evaluation should be compared with the tests developed by non-trained instructors. By analysing their mistakes in details, more functional programs can be designed for measurement training. Secondly, the impact of preservice education and inservice training in measurement and evaluation should be investigated. We need to know which one is more effective. Finally, future studies should explore whether teachers at various school levels and trainers of diverse areas of the corporate world should be trained differently in measurement and evaluation.

References

AFT, NCM, & NEA. (1990). *Standards for teacher competence in educational assessment of students.* Retrieved July 20, 2014 from http://buros.org/standards-teacher-competence-educational-assessment-students

Borich, G. D. (1996). *Effective teaching methods* (3rd ed.). Englewood Cliffs, NJ: Merrill/Prentice Hall.

Cangelosi, J. S. (1990). *Designing tests for evaluating student achievement*. New York: Longman.

FAO. (2014). *Guidelines for trainers*. Retrieved July 20, 2014 from http://www.fao.org/ docrep/t0690e/t0690e0e.htm

Griswold, P. A. (1993). Beliefs and inferences about grading elicited from student performance sketches. *Educational Assessment, 1*(4), 311-328.

Hills, J. P. (1991). Apathy concerning grading and testing. *Phi Delta Kappa, 72*(7), 540-545.

Kennedy, E. (1993). Evaluation of classroom assessment practices: Practitioner criteria. *College Student Journal, 27*, 342–345.

Mehrens, W. A. & Lehmann, I. J. (1984). *Measurement and evaluation in education and psychology* (3rd ed.). New York: Holt, Rinehart and Winston.

O'Sullivan, R. G., & Chalnick, M. K. (1991). Measurement-related course work requirements for teacher certification and recertification. *Educational Measurement: Issues and Practices, 10*(1), 17-19.

Plake, B. S. (1993). Teacher assessment literacy: Teachers' competencies in the educational assessment of students. *Mid-Western Educational Researcher, 6*(1), 21–27.

Simsek, A. (2013). *Ogretim tasarimi* [Instructional design]. Ankara: Nobel.

Stiggins, R. J. (1988). *Revitalizing classroom assessment: The highest instructional priority.* Phi Delta Kappan, 69, 363-368.

Stiggins, R. J. & Conklin, N. F. (1992). *In teachers' hands: Investigating the practices of classroom assessment*. Albany: State University of New York Press.

Stiggins, R. J., Frisbie, R. J., & Griswold, P. A. (1989). Inside high school grading practices: Building a research agenda. *Educational Measurement: Issues and Practice 8*(2), 5-14.

Wiggins, G. (1992). Creating test worth taking. *Educational Leadership, 49*(8), 26-34.

Wise, Wise, S. L., Lukin, L. E., & Roos, L. L. (1991). Teacher beliefs about training in testing and measurement. *Journal of Teacher Education, 42*(1), 37–42.

Zhang, Z. & Burry-Stock, J. A. (2003). Classroom assessment practices and teachers' self-perceived assessment skills. *Applied Measurement in Education, 16*(4), 323-342.