Comparing standard toughness through weighted and unweighted scores by three standard setting procedures

Tsai-Wei Huang
National Chiayi University, Taiwan

Ayres G. D'Costa
The Ohio State University

Abstract

The main purpose of the study was to examine the standard toughness on a weighted and unweighted scoring system by three standard setting procedures: the Nedelsky, Modified Nedelsky, and Percentage (70%) standards-setting methods. There were 174 examinees participating in a 100-item, five-scale test. The Nedelsky codes (1 to 4 representing from the least acceptable option to the correct answer) and the scoring weights (0 to 4 for an option representing partial knowledge from the least to correct) were judged by a panelist of eight raters. Options with weight of 0 were also coded as 0 in the Nedelsky codes. Findings showed that the Modified Nedelsky method appears to be the toughest procedure, while the Percentage method seemed to be the most lenient, especially in the weighted scoring system. The weighted scoring system displayed tougher standards than the unweighted did under both the Nedelsky and Modified Nedelsky procedures, but vice versa under the Percentage procedure. Findings also showed that extremely strict criteria occurred on two scales and the total test, in which no examinees met the cut scores, no matter how the standard-setting methods or the scoring systems changed.

Key words: Standard Setting, Nedelsky, Modified Nedelsky, Percentage, Weighted Scoring System.

Introduction

A standard is known as a cutoff score, a passing score, a mastery score, or a criterion score, all of which can be regarded as a reference score beyond and below which satisfactory and unsatisfactory groups are identified. Standard setting is important because it determines examinees' proficiency levels and influences placement decisions. This is especially significant in high stake tests and license or certification examinations. A tough standard setting would make examinees seem to fail in a test so that very few examinees are selected, while lenient standards would fail to make a distinction among selected candidates.

The standard scores separating levels of proficiency answer the question "How much is enough?" (Zieky & Livingston, 1977, p.2). Beyond the one-point standard score, examinees' scores imply that these examinees are good enough to be proficient. On the multiple-standard-score setting, it means that examinees are good enough to be, for example, proficient but not good enough to be advanced when their scores are between the requirements of the Proficient group and the Advanced group. Therefore, examinees who score at or above a particular cutoff score may be regarded as having a good enough standing in the corresponding proficiency group, while others who score below the standard may be classified at a lower level of proficiency.

Several standard-setting procedures had been developed in literature, e.g., Nedelsky (1954), Angoff (1971), and Ebel (1972). Among them, although the Nedelsky procedure received most criticisms, it was still cited by so many comparative articles (e.g., Behuniak, Archambault, & Gable, 1982; Brennan & Lockwood, 1980; Chang, van der Linden, & Vos, 2004; Cross, Impara, Frary, & Jaeger, 1984; Halpin, Sigmon, & Halpin, 1983; Plake & Kane, 1991; Smith & Smith, 1988; Subkoviak, Kane, & Duncan, 2002; Violato, Marini, & Lee, 2003). As Chang (1999) indicated, about 80% of 40 comparative studies referred to the Nedelsky procedure and showed its lenience. This problem might be due to the fact that the Nedelsky technique allows guessing by examinees and assumes that the remaining distractors are weighted by the same value of 1 (Gross, 1985). Gross indicated this would lead to chance values of the minimum pass index (MPI) and might create a severity bias when MPIs of 1.00 occur for clusters of items. Furthermore, Smith and Gross (1997) had provided a modified Nedelsky method that would overcome these shortcomings, but few empirical data showed its toughness comparisons with the original Nedelsky procedure.

Intuitionally, the percentage method of setting a proper percentage of a total score as a cutoff score was also commonly used (Zieky, 1989). Although this method received few attentions for its easiness and roughness, it was also interesting to explore the value of this intuitional standard-setting procedure. On the other hand, few articles also discussed the effect of weights on the toughness of cutoff score, even though the weights might influence the effects of standard-setting procedures on their toughness. Thus, the main purpose of this study was to examine the interaction effects on standard toughness through the weighted and unweighted scoring systems (SS) and the three standard setting procedures (SP): the Nedelsky procedure (NE), the Modified Nedelsky procedure (MN), and the common-sense 70 percentage procedures (PC70).

Standard Setting Procedures with Weighted Scores

Essentially, the Nedelsky method requires training judges to conceptualize examinee

competency (Chang, 1999). This might involve several steps (Nedelsky, 1954; Zieky, 1989). First, raters need to judge each of the response alternatives in each item to identify the alternatives that a minimally competent examinee would most probably eliminate. Second, after eliminating the alternatives, it is assumed that a minimally competent examinee would guess the correct answer from among the remaining distractors and the correct answer with equal probability. Thus, the probability of hitting the correct answer is equal to the reciprocal of the number of remaining options. Terminologically, this probability is just the MPI of an item. Finally, the raw cut score for the scales or total test is the sum of the MPI's for all of the corresponding items. To obtain a weighted cut score, the scores weighted on the remaining options in each item are multiplied by the probability calculated above and summed.

A major difference between the original and the modified Nedelsky methods is the setting of weights for the remaining distractors. Instead of setting a weight of one for all remaining distractors (including the correct answer), the modified Nedelsky method assumes a weight of 2 for the correct answer, a weight of 1 for the plausible distractors, and a weight of 0 for the least acceptable options. The MPI of each item is computed by the following formula (Gross, 1985; Smith & Gross, 1997):

$$MPI_i = \frac{W_C}{\sum W_i} - \frac{1}{k(\sum W_i)}, \quad (1)$$

where W_C, the weight of the correct option is equal to 2 and $\sum W_i$ is the sum of all option weights. k is a constant that adjusts the minimum and maximal MPI values and is suggested as 5 by Gross (1985) to yielded a minimum MPI of .30 and a maximal of .90.

Finally, the percentage method generally set a 70 percent of the maximal scores for a test is commonly used to determine a raw cut score in an unweighted scoring system. Straightly, in a weighted scoring system, the cut score is determined by the maximal weighted score multiplied by .70.

Method

The data used in this study were collected from the certification examination for Hearing Aid Specialists (D'Costa, 1991). The test consisted of 100 items and contained five scales, including Elicit Patient/Client Hearing History and Problem (EHH), Assess Hearing (ASH), Fit Hearing Aid (FHA), Educate Patient/Client and Family (EPF), and Maintain Professional Standards and Office (MPS). There were 174 examinees who participated in the test.

Two scoring systems, unweighted (USS) and weighted (WSS) on options for each item, were implemented in the calculation of scores. The unweighted system was dichotomous, 1 for the correct answer and 0 for wrong answer, while the weighted scores on options based on the judgments of eight panelists were scribed from 0 to 4 points according to their correctness levels. The eight raters also judged the options by Nedelsky codes from 1 to 4, where 1 was for the least acceptable option(s) and 4 for the correct answer. Between them, the distractors were coded as 2 or 3 based on the accuracy levels judged by these raters. However, an option with weight of 0 was also coded as 0 in the Nedelsky procedure for the same reason of no accuracy as the Nedelsky code 1. Both weighted and unweighted scores were obtained through the WTS program (D'Costa, 1999). Table 1 displayed the means and standard deviations of unweighted

and weighted scores across the five subscales and the whole test.

Table 1 - Means and standard deviations for unweighted and weighted scores by scale

Scales	EHH (15)	ASH (25)	FHA (25)	EPF (20)	MPS (15)	TOTAL (100)
			Unweighted (USS)			
M	9.99	15.47	13.64	13.67	8.18	61.04
SD	2.63	4.01	3.74	3.15	1.95	12.58
MIN	2.00	4.00	5.00	3.00	2.00	26.00
MAX	15.00	24.00	22.00	20.00	13.00	83.00
			Weighted (WSS)			
M	48.41	77.72	66.87	63.10	37.65	293.75
SD	6.40	8.34	5.96	6.02	2.98	22.44
MIN	28.00	57.00	49.00	43.00	29.00	228.00
MAX	59.00	94.00	79.00	74.00	44.00	332.00

Note. Values in the parentheses represent the item numbers for each scale and the total test.

MPI illustration

To illustrate the calculation of MPIs in scoring systems, Table 2 shows MPIs calculated by the three standard-setting procedures for scale EHH (Elicit Hearing History). First, as can be seen, the correct answer of the first item was located at the second option with a weight of 4 and also a Nedelsky code of 4. Other options earned the same weight of 2 from the raters, but the last three distractors were regarded as the options that the least able persons would not choose. In this case, only the first and second options remained. Thus, an examinee with the least ability could guess the correct answer from the remaining options with equal probability on each option, i.e., $p = .5$. That is, the first item earned a MPI value of 0.5 in the unweighted-scoring situation. Multiplying the sum of weights for the first two options by the MPI will obtain a cut score of 3.00 for the first item in the weighted-scoring situation by the Nedelsky procedure. Similarly, through the equation (1), the MPI of the first item for the Modified Nedelsky Procedure would be calculated as 0.6 and 3.6 in the unweighted- and weighted- scoring situations, respectively. Finally, it would be much easier to calculate the MPI's for the Percentage procedure. They are just the values of .7's in the unweighted-scoring situation, and the products of a weight of 4 on

the correct option with the .7 in the weighted-scoring situation. As regards the cut scores, based on different situations, it is easy to just add the MPI of each item for all scales and the total test, respectively.

Note that in some items there were some zeros appearing both in the scoring weight and the Nedelsky code. This is based on the logic that a distractor without any partial knowledge (weighted as 0) would be abandoned by an examinee absolutely.

Table 2
The calculation of MPI by NE, MN, and PC70 Methods for The EHH scale

Item	Scoring Weights	Nedelsky Codes	Unweighted (USS)			Weighted (WSS)		
			NE	MN	PC70	NE	MN	PC70
001	24222	24111	0.50	0.60	0.70	3.00	3.60	2.80
002	24220	24110	0.50	0.60	0.70	3.00	3.60	2.80
003	11341	11241	0.50	0.60	0.70	3.50	4.20	2.80
005	14110	14110	1.00	0.90	0.70	4.00	3.60	2.80
006	42220	41110	1.00	0.90	0.70	4.00	3.60	2.80
009	11143	11143	0.50	0.60	0.70	3.50	4.20	2.80
010	14212	14112	0.50	0.60	0.70	3.00	3.60	2.80
012	24111	14111	1.00	0.90	0.70	4.00	3.60	2.80
020	22432	11431	0.50	0.60	0.70	3.50	4.20	2.80
022	24210	24110	0.50	0.60	0.70	3.00	3.60	2.80
024	22224	21114	0.50	0.60	0.70	3.00	3.60	2.80
055	42320	41210	0.50	0.60	0.70	3.50	4.20	2.80
060	11140	11140	1.00	0.90	0.70	4.00	3.60	2.80
070	34211	34111	0.50	0.60	0.70	3.50	3.60	2.80
087	14120	14110	1.00	0.90	0.70	4.00	3.60	2.80
		Total	10.00	10.50	10.50	52.50	57.00	42.00

Cutoff scores

After summing the calculated MPIs, cut scores under different standard setting procedures (SP) and scoring systems (SS) for each scale and the whole test were shown in Table 3. As can be seen in the unweighted situation, the cutoff scores seemed to present a consistent order pattern across scales in which scores by the NE procedure were greater than those by the MN procedure and than those by the PC70 procedure except for cutoff scores in the EHH scale (EHH). This might imply the strong magnitudes of toughness of the NE procedure and the strong magnitudes of leniency of the PC70 procedure. Yet, a few changes of order occurred in the weighted scoring system. Contrary to its previous medium role under the raw data, the MN procedure played a severe role in the EHH scale and the EPF scale. The PC70 procedure still stood at a lenient state across scales. Interestingly, regarding the whole test, the NE procedure was the most lenient under the unweighted situation but presented the highest threshold under the weighted situation.

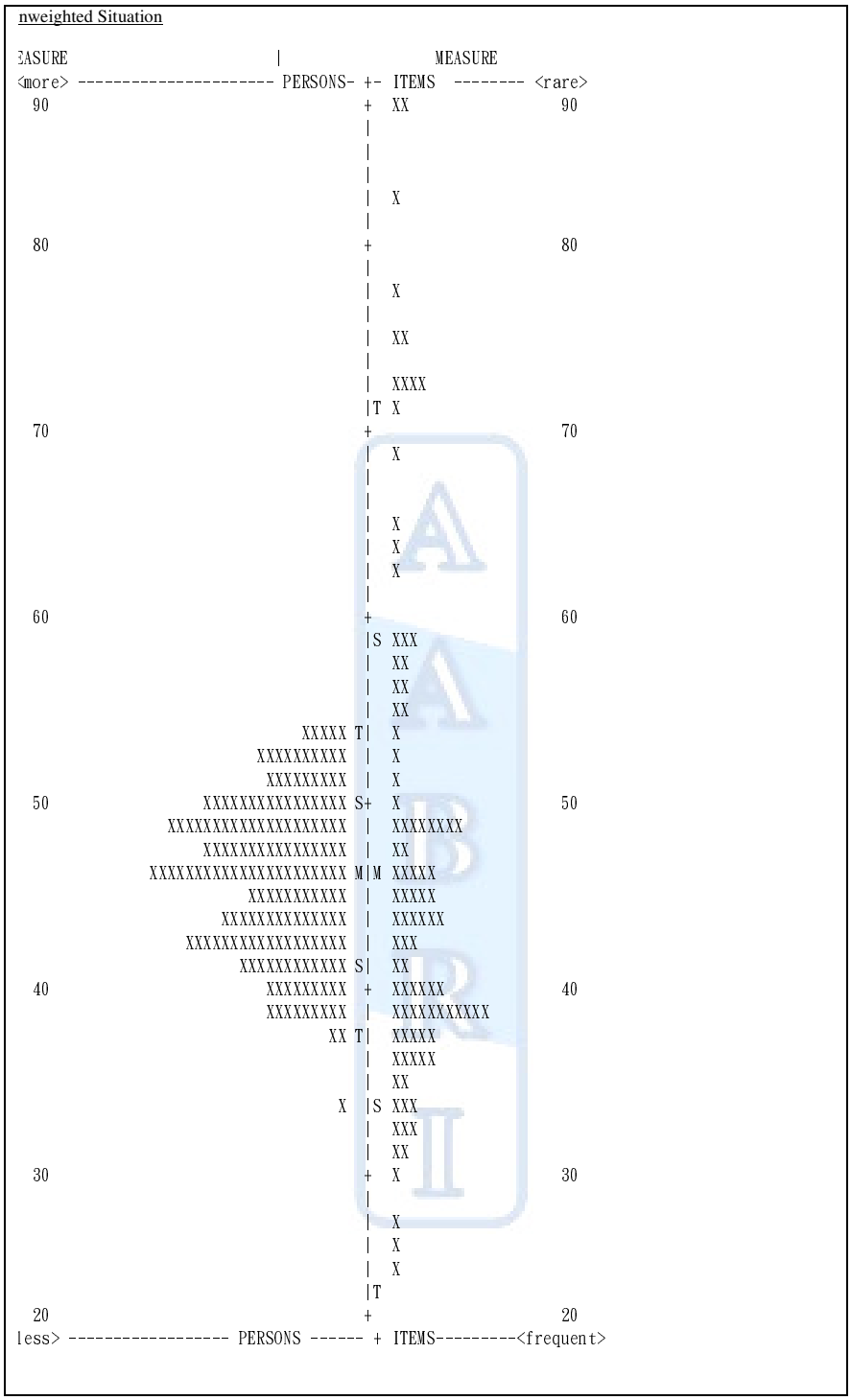
In contrast, the MN procedure never obtained the highest cut score in any scale in the unweighted scale, yet it was the toughest one for the whole test. The PC70 procedure seemed to play a consistent role in the scales.

Table 3
Cutoff scores by SP and SS

	Number	Unweighted (USS)			Weighted (WSS)		
Scale	of Items	NE	MN	PC70	NE	MN	PC70
EHH	15	10.00	10.50	10.50	52.50	57.00	42.00
ASH	25	23.00	21.30	17.50	97.50	91.80	70.00
FHA	25	21.50	21.40	17.50	96.00	93.60	70.00
EPF	20	16.00	15.60	14.00	72.50	72.60	56.00
MPS	15	13.00	12.30	10.50	83.50	80.10	42.00
Whole	100	57.00	80.10	70.00	375.50	370.20	280.00

Result and Discussion- Data map and Description

A person-item map (Linacre & Wright, 2000) shown in Figure 1 for unweighted and weighted total logit scores provided an overall understanding of person ability and item difficulty. As can be seen in unweighted situation, the mean logit for persons was 47 logits, almost the same as that for items (47 logits), but the standard deviation of persons was smaller than that of items. Most person measures were distributed from 35 logits to 55 logits, a narrower range than the range in which the item difficulties were dispersed so that examinees could not solve those items whose difficulty was beyond one standard deviation.


nweighted Situation
EASURE | MEASURE
<more> ---------- PERSONS- +- ITEMS -------- <rare>
90 + XX 90
80 + 80
70 + 70
60 + 60
50 + 50
40 + 40
30 + 30
20 + 20
less> ---------- PERSONS ------ + ITEMS---------<frequent>

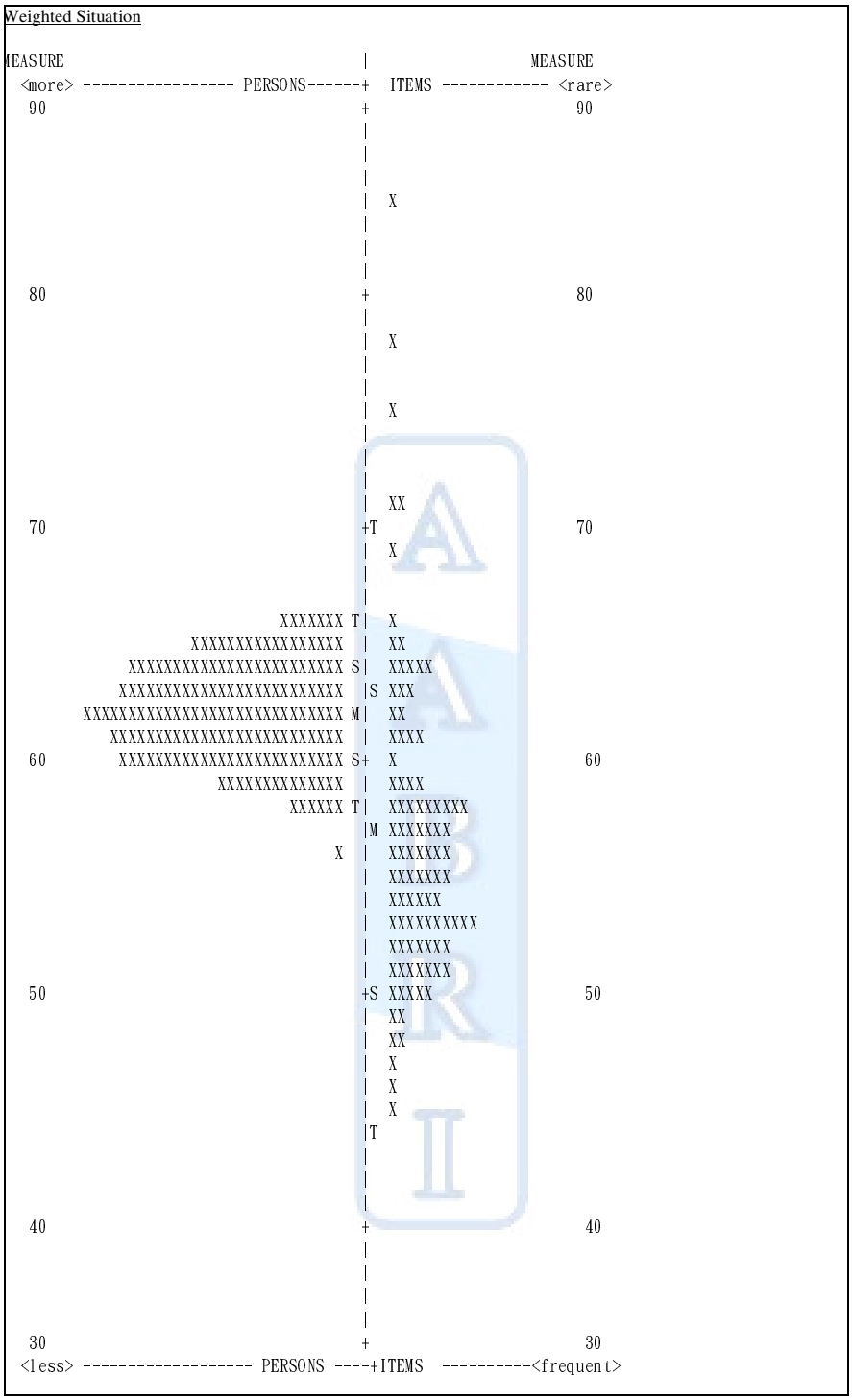


Figure 1. Maps of persons and items in scores unweighted and weighted situation

On the other hand, the mean logit for person was almost 62 logits. Higher than that for item (mean closer to 57) in the weighted situation, but the standard deviation of persons was smaller than that of items again. Most person measures were distributed approximately from 58 logits to 65 logits, again within a narrower range than the range in which the item difficulties

were dispersed. Interestingly, the dispersion of weighted total logit scores revealed a lower level of dispersion than that of unweighted ones. This much concentration indicates that examinees seemed to have the capability of overcoming most items.

In addition, Table 1 also provides a summary of examinees' performances. The total raw scores (unweighted) ranging from 26 to 83 in the 100-item test with a mean of score 61.04 and standard deviation of 12.58 was close to the upper bound and revealed a slight negatively skewed distribution. On the other hand, total weighted scores ranged from 228 to 332 (*M = 294.33, SD = 22.44).* Although the enlarged values of mean and standard deviation in the weighted situation are reasonable, the ratio of mean to standard deviation still inflated (from 4.85 to 13.09). This implied a more condensed data pattern in weighted scores.

Passing Rates

Based on these cutoff scores, the numbers of passed examinees were counted in Table 4. As can be seen, passing frequencies affected by the factors of SP and SS varied across EHH, ASH, and EPF scales, but did not much vary in the FHA and MPS scales and the whole test. The passing rates of the NE and MN standard-setting procedures showed a similar trend across scales both in the unweighted scoring system and the weighted scoring system. But the PC70 procedure exhibited higher passing rates than the other two procedures in both scoring systems. On the other hand, the USS exhibited consistently higher passing rates than the WSS system did in both the NE and MN procedures, but with contrasting results shown in the PC70 procedure.

Table 4
Numbers of examinees passing the cutoff scores by SP and SS

	NE	MN	PC70	N	NE	MN	PC70	N
	EHH				ASH			
Unweighted	90 (22.33)	54 (13.40)	54 (13.40)	198 (49.13)	0 (0.00)	2 (1.03)	40 (20.62)	42 (21.65)
Weighted	46 (11.41)	4 (0.01)	155 (38.46)	205 (50.87)	0 (0.00)	2 (1.03)	150 (77.32)	152 (78.35)
N	136 (33.75)	58 (14.39)	209 (51.86)	403	0 (0.00)	4 (2.06)	190 (97.94)	194
	FHA				EPF			
Unweighted	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	30 (9.77)	30 (9.77)	64 (20.85)	124 (40.39)
Weighted	0 (0.00)	0 (0.00)	54 (0.00)	54 (100.00)	15 (4.89)	15 (4.89)	153 (49.84)	183 (59.61)
N	0 (0.00)	0 (0.00)	54 (100.00)	54	45 (14.66)	45 (14.66)	217 (70.68)	307
	MPS				Whole test			
Unweighted	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
Weighted	0 (0.00)	0 (0.00)	20 (100.00)	20 (100.00)	0 (0.00)	0 (0.00)	123 (100.00)	123 (100.00)
N	0 (0.00)	0 (0.00)	20 (100.00)	20	0 (0.00)	0 (0.00)	123 (100.00)	123

Note. Values in the parentheses represent the passing rates (%) for each scale and the whole test.

For further analysis, since there were no Chi-square values calculated in the FHA and MPS scales and the whole test for only one non-zero cell frequency existed in the corresponding contingency tables, no comparisons occur. For the ASH scale, since no interaction effect existed, the main effects of SP and SS on passing rates were found ($\chi^2_{1,\underline{N}=194} = 178.33$, $p < .001$, 1 degree of freedom because zero frequency occurred in NE cell, and $\chi^2_{1,\underline{N}=194} = 62.37$, $p < .001$, respectively). Figure 2 reveals the Main effects of SP and SS factors for the ASH scale, respectively. As can be seen, the PC70 procedure displayed larger effect then the NE and MN procedures did, no matter in which scoring system. On the other hand, the WSS dominated the USS on the PC70 standard setting procedure, but had almost no influence on the other two standard setting procedures. Next, we will examine the interaction effects for the EHH scale and the SPF scale.

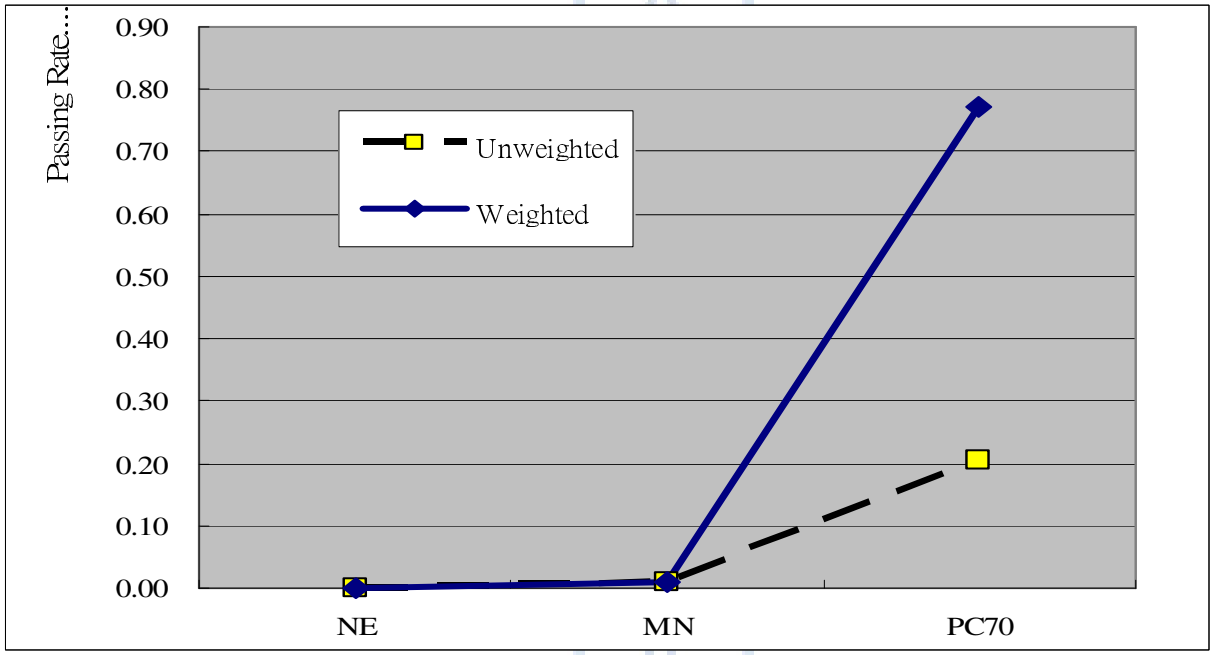


Figure 2. Main effects of scoring systems and standard setting procedures on the ASH scale

Interaction effect by the factors of SP and SS

The interaction effects of the SP and the SS factors on passing rates were only revealed on the EHH scale and the SPF scale ($\chi^2_{2,\underline{N}=403} = 106.06$, $p < .001$ *and* $\chi^2_{2,\underline{N}=307} = 36.51$, $p < .001$, respectively). That is, the effect of SP will be influenced by the factor of SS, and vice versa. Further simple main effect analyses in the EHH scale are shown in Figure 3. As can be seen in Figure 3 (EHH scale), passing rates under the unweighted scores were higher than under the weighted scores for both the NE and MN procedures, but vice versa for the PC70 procedure. A similar situation for and the EPF scale is seen in Figure 4, but the differences of passing rates between the two score systems are smaller than those in the EHH scale. Further simple main effects of SP and SS on the EHH and EPF scales are explored in next two paragraphs.

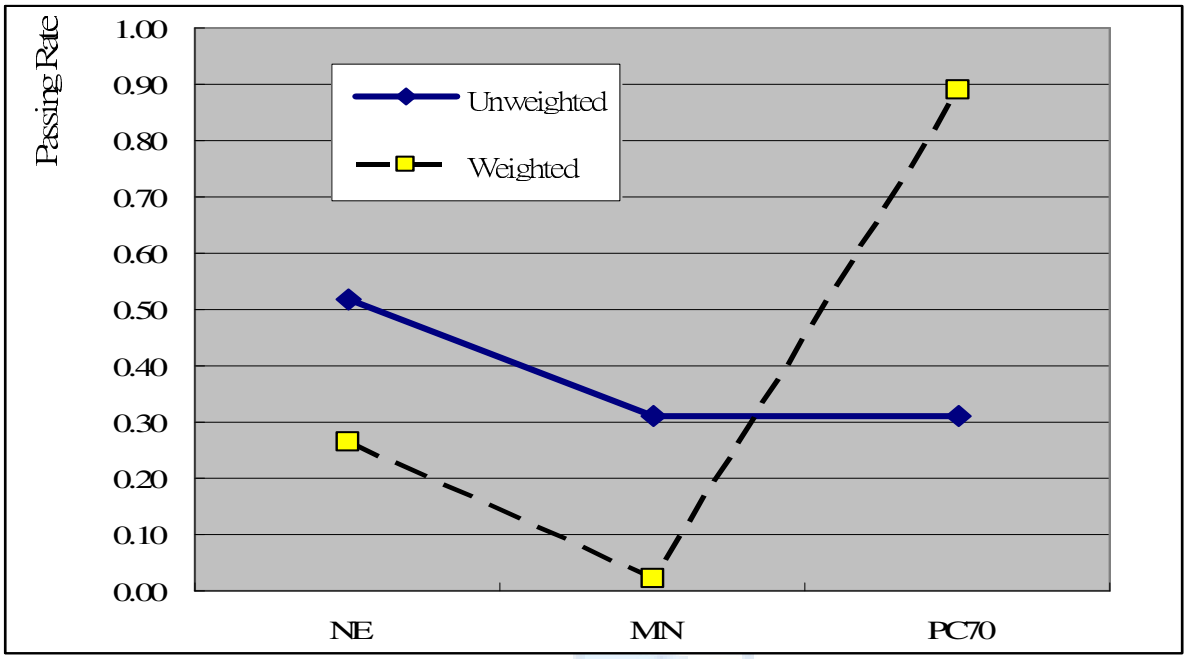


Figure 3. Interaction effects by factors of SP and SS in the EHH scale

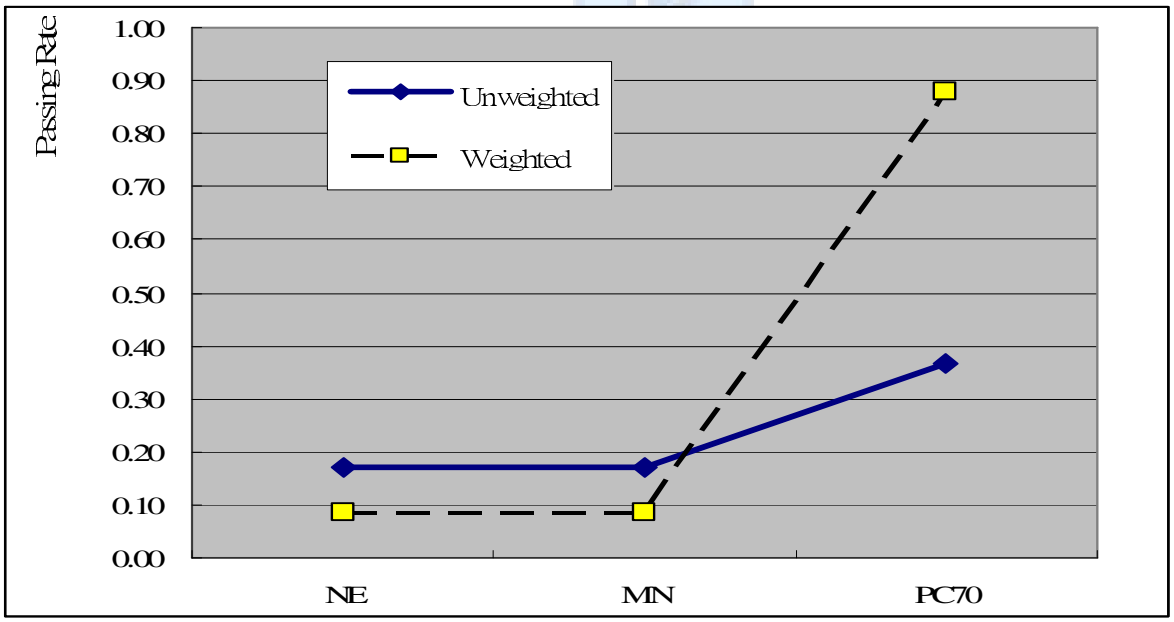


Figure 4. Interaction effects by factors of SP and SS in the EPF scale

Simple main effects for SP

In the EHH scale, the simple main effects of standard setting procedures on passing rates showed significant differences under two scoring systems ($\chi^2_{(2,N=198)} = 13.09$, $p < .01$ *for* USS *and* $\chi^2_{(2,N=205)} = 177.79$, $p < .001$ *for* WSS, respectively*)*. For further comparisons in the unweighted situation, the NE procedure dominated other procedures (both *Z*s equal to 3.92, *p*< .01), but the comparison between the MN and PC70 procedures were not significantly different. For the weighted scores, the number of passing examinees calculated by the PC70 procedure appeared significantly larger than those by other two procedures (*Z=11.83, p* < .001 and *Z=16.25, p* < .001, compared with the NE and MN procedures, respectively). Note that, the NE procedure still dominated the MN procedure (*Z=6.42, p* < .001).

In the EPF scale, similar results showed that the simple main effects of standard setting procedures on passing rates significantly differed across the two scoring systems ($\chi^2_{(2,N=124)} = 18.65$, $p < .001$ *for* USS *and* $\chi^2_{(2,N=183)} = 265.01$, $p < .001$ *for* WSS, respectively*)*. Through post hoc comparisons, it was found that no matter for the unweighted or weighted scores, the PC70 procedure possessed the highest rate, and was the most lenient one (both comparative *Z*s equal to 14.80, *p <.001*), but in this scale, the ME and the MN procedures were not different from each other.

Simple main effects for SS

In the EHH scale, the simple main effects of scoring systems on passing rates showed significant differences under the NE procedure ($\chi^2_{(1,N=136)} = 14.24$, $p < .001$*)*, the MN procedure ($\chi^2_{(1,N=58)} = 343.10$, $p < .001$*)*, and the PC70 procedure ($\chi^2_{(1,N=209)} = 48.81$, $p < .001$*)*. After post hoc comparisons, it was found that the unweighted system exhibited higher passing rates than the weighted system did in both the NE and MN procedures (both comparative *Z*s equal to 4.83, *p <.001*), but vice versa in the PC70 procedure (*Z= -11.50, p* < .001).

In the EPF scale, the simple main effects of scoring systems on passing rates showed significant differences under the NE procedure ($\chi^2_{(1,N=45)} = 5.00$, $p < .05$*)*, the MN procedure ($\chi^2_{(1,N=45)} = 5.00$, $p < .05$*)*, and the PC70 procedure ($\chi^2_{(1,N=247)} = 57.33$, $p < .001$*)*. Further comparisons showed similar results in the EHH scale, both comparative *Z*s in the NE and MN procedures equal to 4.83 (*p <.001)*, but a *Z* value equal to -9.85 *(* $p < .001$) in the PC70 procedure.

Conclusions

This study examined the interaction effects of three standard setting procedures (Nedelsky, Modified Nedelsky, and Percentage) and two scoring systems (unweighted and weighted) on standard toughness for the Hearing Aid Specialists test. Findings showed that individual main effects of standard setting procedures and scoring systems existed in the ASH scale (Assess Hearing), in which more participants passed the cutoff scores in the Percentage procedure and in the weighted scoring system than in other procedures and the unweighted scoring system, respectively. The interaction effects of standard setting procedure and scoring

system existed in the EHH scale (Elicit Patient/Client Hearing History and Problem) and the EPF scale (Educate Patient/Client and Family). Through simple main effect and post hoc analyses, findings showed inconsistent levels of toughness of standard setting procedure across the EHH and EPF scales in the unweighted situation, but consistent results in the weighted situation in which the Percentage procedure performed more leniently than the Nedelsky and the Modified Nedelsky procedures. Moreover, the Modified Nedelsky method appears to be the toughest one; few examinees met the criterion scores established. On the other hand, the simple main effects of the weighted scoring system displayed tougher standards than the unweighted one did under both the Nedelsky and Modified Nedelsky procedures, but vice versa under the Percentage procedure.

Findings also showed that extremely strict criteria occurred on the scales FHA, MPS and the whole test, in which no examinees could meet the cut scores, no matter how the standard-setting methods or the scoring systems changed. This seemed to indicate the limitations of these applied standard-setting methods and scoring systems. For future studies, feasible standard-setting approaches, such as Angoff or Bookmark et al., may be considered to cooperate with scoring systems in the setting of cutoff scores.

References

Angoff, W. (1971). Scales, norms and equivalent scores, In R. L. Thorndike (Ed.), *Educational Measurement*. Washington, D.C.: American Council on Education.

Behuniak, P., Archambault, F. X., & Gable, R. K. (1982). Angoff and Nedelsky standard setting procedures: implications for the validity of proficiency test score interpretation. *Educational and Psychological Measurement, 42*, 247-801.

Brennan, R. L., & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using Generalizability theory. *Applied Psychological Measurement, 4*(2), 219-240.

Chang, L. (1999). Judgmental item analysis of the Nedelsky and the Angoff standard-setting methods. *Applied Measurement in Education, 12* (2), 151-165.

Chang, L. van der Linden, W. J., & Vos, H. J. (2004). Setting standards and detecting intrajudge inconsistency using interdependent evaluation of response alternatives. *Educational and Psychological Measurement, 64*, 781-801.

Cross, L. H., Impara, J. C., Frary, R. B., & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the National Teacher Examinations. *Journal of Educational Measurement, 21,* 113-129.

D'Costa, A. (1991). *1991 Role delineation study hearing aid specialists.* Livonia, MI: National Board for Certification in Hearing Instrument Sciences (NBC-HIS).

D'Costa, A. (1999). *WTSCOR: Creates Weighted and Unweighted Scored Response Files* (Version 2.0) [Computer programming language]. Unpublished computer program.

Ebel, R. L. (1972). *Essentials of education measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Gross, L. J. (1985). Setting cutoff scores on credentialing examinations: A refinement in the Nedelsky procedure. *Evaluation and the Health Professions, 8,* 469-493.

Halpin, G., Sigmon, G., & Halpin, G. (1983). Minimum competency standards set by three divergent groups of raters using three judgmental procedures: Implications for validity. *Educational and Psychological Measurement, 43,* 185-196.

Linacre, J. M., & Wright, B. D. (2000). *A user's guide to WINSTEPS: Rasch-model computer program*. Chicago: MESA Press.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14,* 3-19.

Plake, B. S., & Kane, M. T. (1991). Comparison of methods for combining the minimum passing levels for individual items into a passing score for a test. *Journal of Educational Measurement, 28*, 249-256.

Smith, R. M., & Gross, L. J. (1997). Validating standard setting with a modified Nedelsky procedure through common item test equating. *Journal of Outcome Measurement, 1*(2), 164-172.

Smith, R. L., & Smith, J. K. (1988). Differential use of item information by judges using Angoff and Nedelsky procedures. *Journal of Educational Measurement, 25*, 259-274.

Subkoviak, M. J., Kane, M. T., & Duncan, P. H. (2002). A comparative study of the Angoff and Nedelsky methods: Implications for validity. *Mid-Western Educational Researcher, 15*, 3-7.

Violato, C., Marini, A., & Lee, C. (2003). A validity study of expert judgment procedures for setting cutoff scores on high-stakes credentialing examinations using cluster analysis. *Evaluation and the Health Professions, 26*, 59-73.

Zieky, M. J. (1989). Methods of setting standards of performance on criterion referenced tests. *Studies in Educational Evaluation, 15,* 335-338.

Zieky, M. J., & Livingston, S. A. (1977). *Basic skills assessment manual for setting standards*. Princeton, NJ: Educational Testing Service.