Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 19, Number 14, November 2014 ISSN 1531-7714

Editorial Changes and Item Performance: Implications for Calibration and Pretesting

Heather Stoffel, Mark R. Raymond, S. Deniz Bucak, Steven A. Haist

National Board of Medical Examiners

Previous research on the impact of text and formatting changes on test-item performance has produced mixed results. This matter is important because it is generally acknowledged that *any* change to an item requires that it be recalibrated. The present study investigated the effects of seven classes of stylistic changes on item difficulty, discrimination, and response time for a subset of 65 items that make up a standardized test for physician licensure completed by 31,918 examinees in 2012. One of two versions of each item (original or revised) was randomly assigned to examinees such that each examinee saw only two experimental items, with each item being administered to approximately 480 examinees. The stylistic changes had little or no effect on item difficulty or discrimination; however, one class of edits – changing an item from an open lead-in (incomplete statement) to a closed lead-in (direct question) – did result in slightly longer response times. Data for nonnative speakers of English were analyzed separately with nearly identical results. These findings have implications for the conventional practice of repretesting (or recalibrating) items that have been subjected to minor editorial changes.

A fundamental assumption of equating and calibration is that the text and layout of any item designated as an equator, linking item, or anchor item must remain constant across test forms. Psychometricians counsel their clients to follow a simple but important rule: if an item changes, it is a new item, and it cannot be designated as a common item for scaling, equating, or calibration. Testing agencies are encouraged to apply this rule to any type of revision, ranging from minor edits to more extensive changes (Kolen & Brennan, 2004). This rule is particularly relevant in recent years given the emphasis on measuring student growth in K–12 education (Young, 2006) and on progress testing in higher education (Schaap, Schmidt, & Verkoeijen, 2012), both of which assume that test content remains unchanged over test administrations.

While this rule certainly seems prudent, it is also costly as it displaces the capacity to pretest original test items and increases the time required for an item to become operational. The practical challenge is that test materials require frequent revision – perhaps more now than in the past – because of continual changes in knowledge, technology, and social convention. Revisions are prompted in some instances by changes in authoritative style guidelines (e.g., *The Chicago Manual of Style*). In other instances, changes in the medical lexicon, such as the renaming of microorganisms or medical disorders, stimulate revision. It is now common, for example, to refer to "human immunodeficiency virus infection" as simply "HIV infection" in medical text. In addition, changing technology is often the impetus for change. For example, the replacement of traditional medical x-rays with digital imaging required that the word film be removed from test questions that contained the phrase x-ray film. In each of these instances, the first decision for a test developer is whether to continue with the old style or update hundreds or even thousands of test questions to reflect the new style. Assuming that the change is desirable, one then must decide whether repretesting or recalibration is necessary.

On one hand, consistent application of the "revised item means new item" rule is judicious because test items often perform in unpredictable ways and even minor changes in terminology, option order, item order, and text formatting have been shown to impact item statistics (Brennan, 1992; Cizek, 1994). On the other hand, it seems intuitive to expect that minor changes in punctuation, style, or word choice will have minimal influence on item performance. However, intuitions can be misleading and there is insufficient documentation to guide practitioners when making decisions about the need to repretest.

Prior research has shown that revising test items by adding or removing information does have an effect on item difficulty and discrimination (Case, Swanson, & Becker, 1996). Also, substituting medical terminology with lay terminology affects performance, but differentially depending on examinee characteristics such as ability and native language (Eva, Brooks, & Norman, 2001; Norman, Arfai, Gupta, Brooks, & Eva, 2003). In a similar vein, studies of alterations in language complexity have demonstrated an effect on item statistics (Abedi, in press; Cassels & Johnstone, 1984; Plake & Huntley, 1984), albeit sometimes the effects are small (Bornstein & Chamberlain, 1970; Green, 1984). Because test items are often revised to correct specific types of problems such as negatively-worded stems, convergence among options, and other common flaws, there also has been interest in studying the impact of these types of improvements (Haladyna & Rodriguez, 2013). Indeed, such changes have been shown to impact item difficulty and examinee performance (Caldwell & Pate, 2013; Cassels & Johnstone, 1984; Downing, 2005; Dudycha & Carpenter, 1973; Green, 1984; Tarrant & Ware, 2008; Violato & Marini, 1989). The previously cited studies generally indicate that many types of revisions affect item performance, thereby supporting the need to repretest.

In contrast, a few reports, all of them unpublished, suggest that many types of stylistic edits have little or no impact on the statistical characteristics of test items. O'Neill (1986) found no significant differences in item performance on a pharmacy licensure test for which small changes were made to abbreviations, symbols, or drug names (i.e., generic vs. proprietary). A later study by Webb and Heck (1991) offered further support that stylistic changes had no detectable effect on item difficulty. Most recently, Zhang and Zhu (2013) studied the effect of a small number of minor changes (e.g., updating drug names, editorial or stylistic manipulations) on examinee performance; results demonstrated that these minor changes had little or no impact on item performance. The results of these few unpublished studies suggest that repretesting is not required for many types of edits. While these studies did not always document the specific type of edit, the reports did refer to them as minor stylistic changes, and it seems reasonable to attribute the lack of effect to the minor nature of the edit. However, additional research is needed to confirm or refute these findings, and to make the results more accessible to test developers.

The purpose of this research was to determine the extent to which different types of minor editorial changes affect item performance. This experiment expanded previous research in four ways. First, we included a larger sample of items (65 pairs) than prior studies. Second, to improve statistical power and facilitate generalization, items were categorized according to the class of edit, with most classes consisting of several items. Third, three outcome variables were studied: item difficulty, item discrimination, and time required to respond to each item. All three dependent measures are important because it is possible that certain editorial or stylistic changes could, for example, affect reading time without affecting item difficulty or discrimination. Finally, we studied the effect of editorial changes for a subset of examinees who were not native speakers of English, recognizing that language fluency may moderate the impact of editorial changes (Abedi, *in press*).

Method

Data Source

The test items for this study consisted of 65 pairs of multiple-choice questions (MCQs) appearing on Step 1 of the United States Medical Licensing Examination®, a computer-based examination. The study included 31,918 examinees taking Step 1 for the first time between May 2011 and May 2012; 32% of examinees indicated that they had learned English as a second language (ESL). Each test form consisted of 322 items, with a proportion of these designated as unscored (pretest) items. The 65 pairs of study items were treated as pretest items and did not count towards examinee scores. Each item pair consisted of an original and a revised version, and items were classified into one of seven categories based on the type of edit as indicated in Table 1. Although none of the edits was intended to change the underlying meaning of the item, it is apparent from Table 1 that some of the changes were more extensive than others. For example, closing the lead-in to make a complete question requires adding words to an item, compared with the smaller change of removing an apostrophe *s* from a word. Two categories, *adding clarifying information* and *removal of superfluous information*, tend to have more heterogeneous changes and some may cross the line from minor to major revision.

Experimental items were distributed across test forms and examinees such that each examinee saw a random subset of pretest items from an entire pretest pool of several hundred items. For the present study, each examinee responded to only two of the experimental items chosen at random with the constraint that an examinee would not be administered both an original version and a revised version of the same item. Each item was seen by an average of 481 examinees, with the actual sample size per item ranging from 401 to 561. On average, only about 8% of examinees had any two items in common. Given that most examinees saw a unique two-item set of study items, the administration closely approximated a between-subjects design with each examinee measured on different items. Each item pair can be regarded as a replication across independent samples of examinees, with each replication being on a different scale determined by the content and statistical properties of that item pair.

Table 1. Classes of Editorial Changes

Class of Edit	Abb rev	N	Explanation and/or Examples
Adding clarifying information	ACI	6	Include additional information, sometimes in parentheses. For example, adding BMI to existing height and weight information.
Closed lead-in	CLI	14	Change stem from open ended, where each option completes the stem, to the interrogative form ending with a question mark. The change is typically from a phrase like, "The most likely diagnosis is" to, "Which of the following is the most likely diagnosis?"
Adding text to items with graphics	PIC	8	Rather than just displaying a graphic, change text to explicitly say "in the photograph shown."
Removal of possessives	POS	13	Remove apostrophes from eponyms. For example, "Wilson disease" instead of "Wilson's disease."
Removal of explanatory information	REI	4	Delete information thought to be unnecessary for examinees with this level of training, such as removing the parenthetical abbreviation from "urea nitrogen (BUN)." Another example is to remove parentheses that include the secondary Latin name for a disease.
Removal of superfluous information	RSI	7	Remove information that has become obsolete, such as "film" from "x-ray film."
Replacing term with synonym	SYN	13	Interchange essentially synonymous terms, such as "limbs" with "extremities" or "neonate" with "newborn."

Analyses

Descriptive statistics and inferential tests are reported for item difficulty, discrimination, and response time (RT). Significance testing was done at two levels for each outcome variable as further described below. Data were first aggregated within each of the seven edit classes using statistical procedures for meta-analysis (Hedges & Olkin, 1985;

Lipsey & Wilson, 2000). Mean effect sizes (i.e., change in difficulty, discrimination, and RT) and confidence intervals (CIs) were calculated for each class of edits. The Q statistic, which is distributed as χ^2, was used to evaluate the consistency of findings across replications within each class of edits. A significant Q suggests that the variability in effect sizes cannot be attributed to sampling error and may indicate the presence of some systematic source of variability. The second level of analysis was at the item pair level. Effect sizes and CIs were computed for each of the individual 65 item pairs to determine if there was a change in the outcome variables. CIs that did not include zero were regarded as statistically significant.

Item difficulty. Item means (p values) were obtained and differences in p values were plotted for all item pairs within a class of edits. Next, odds ratios were computed and served as the basis for cumulating results across item pairs and for evaluating statistical significance. Odds ratios have statistical properties that make them more desirable than p values for assessing group differences (Fleiss, 1994). Odds ratios were transformed to their natural logarithm prior to aggregation; log-odds ratios that are significantly different from zero would indicate that the original and revised items within that class of edits are not equally difficult. It is noted in passing that log-odds ratios are comparable to the logit unit of item difficulty under the Rasch model. After evaluating log-odds ratios within each class of edits, odds ratios for each of the 65 item pairs were inspected.

Item Discrimination. The correlation (r) between each item score (0 or 1) and the total score was obtained, and differences in r for the original and revised version were computed and plotted. For purposes of data aggregation and statistical testing, r was subjected to Fisher's Z transformation (Zr), and all differences in correlations were calculated from Zr.

Response time (RT). The time, in seconds, for each examinee to respond to an item was recorded by the test administration software. For descriptive purposes, we report median RTs across examinees for each item. For inferential purposes, RTs were subjected to a logarithmic transformation to compensate for the positive skew they typically exhibit (Ratcliffe, 1993; van der Linden, 2006). Log-transformed RTs were then used as the basis for computing effect sizes and aggregating results across item pairs within each class of edit. The mean difference in log RT for each item pair was also evaluated for statistical significance.

The preceding analyses were first completed for all examinees and then separately for ESL examinees. Item pairs that exhibited large or significant differences were triaged for review to identify the possible source of the differences. Although we completed a large number of statistical tests without adjustment for type I error rate, we felt that a liberal approach to data interpretation was justified in the interest of not overlooking any potentially significant effects associated with making editorial changes.

Results

All Examinees

Item Difficulty. Figure 1 displays the change in p value for individual pairs of items within the seven classes of edits, with positive values indicating that the revised item was easier than the original item. The changes in p values for individual items range from about −0.05 to 0.04. The X's is Figure 1 correspond to the mean for each class of edit. The largest within-class mean difference is for *removal of possessives* (POS), with a mean change of −0.016.

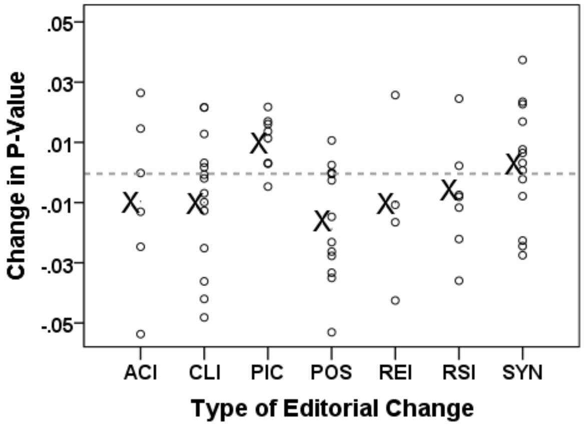

Figure 1. Change in p values for different types of editorial changes.

Table A.1 in the Appendix summarizes the results of the statistical meta-analyses. The log-odds ratio for each item pair was computed and weighted by the inverse of its standard error for cumulating effects across replications within each class of edit. The only type of edit for which the log-odds ratio reached statistical significance was for *removal of possessives*, which barely reached statistical significance. None of the six remaining classes of edits produced significant changes in difficulty. None of the *Q* tests for homogeneity of effect sizes reached statistical significance, suggesting that any variation in changes in item difficulty is likely due to sampling error. For completeness, odds ratios for all 65 individual item pairs were inspected. One item pair within the *closed lead-in* (CLI) category exhibited an odds ratio of 0.621 (CI = 0.395 to 0.978). This item became slightly more difficult, with the *p* value dropping from 0.934 to 0.898. Of note, none of the individual odds ratios within the *removal of possessives* category was significant.

Item Discrimination. Figure 2 shows the change in *r* for each of the item pairs, which tend to be symmetrically distributed around zero within each class. There is considerable variability in differences in *r*, owing partly to the fact that correlation coefficients typically lack stability and have relatively large standard errors. To more formally evaluate these differences, values of *r* were converted to *Zr* as a measure of effect size and combined across all items within each class of edit. There were no significant changes in *Zr* either for class of edit or for the 65 individual item pairs also. Thus, the edits had no discernable impact on item discrimination (see Appendix, Table A.2).

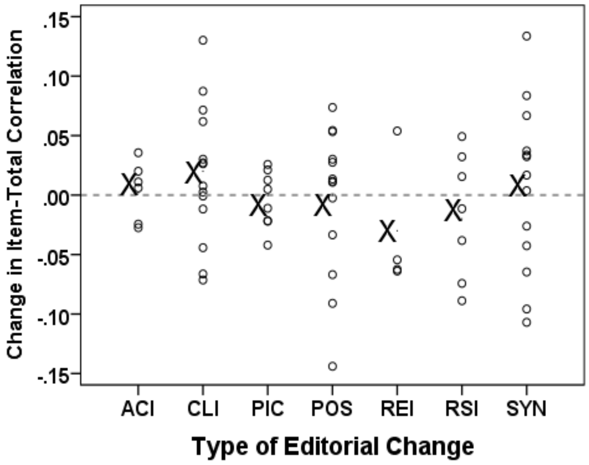

Figure 2. Change in item-total correlation, *r*, for different types of editorial changes.

Response Time. Analyses of RTs mirrored those for item difficulty and discrimination except that medians were used in the graphic summary, while log RTs served as the basis for computing and cumulating effect sizes. Figure 3 shows the change in median RT for the 65 item pairs. The differences ranged from an 8.4-second increase to a 10.4-second decrease, with most changes falling within about ± 5 seconds. Figure 3 does suggest that RTs are slightly longer for *closed lead-in* and possibly for *removal of explanatory information* (REI). As indicated in Table A.3 (see Appendix), statistical testing indicates that the change in RT for *closed lead-in* of 1.4 seconds was significantly different from zero, suggesting that direct questions (revised version) required slightly more time than incomplete statements or open-ended lead-ins (original version). Also, the *Q*-test test for homogeneity of effect sizes for *removal of explanatory information* was statistically significant, Q (3 df) = 7.98, $p < 0.05$, indicating that variability in log RT effect sizes for that class of edits could not be explained by sampling error alone. Of the four item pairs in this class, one item took 5.2 seconds longer, while the other three items had changes in RTs of 2, 2, −1.2, and −1.7. These differences, more fully discussed below, raise the possibility that differences in RT might vary according to the specific type of information removed. Of the 65 individual item pairs, six were found to have changes in RT significantly different from zero. Three of the significant changes were in the *closed lead-in* category, all of which required longer response times (3.0, 8.3, and 8.4 seconds); these differences are consistent with the RT results reported above for the entire class of *closed lead-ins*. Longer RTs were also required for one item pair belonging to the *removal of possessives* category (5.1 seconds longer), and for one item pair in the *adding text to items with graphics* (PIC) category (6.9 seconds). There was one item for which *adding clarifying information* (ACI) resulted in a faster RT (−3.1 seconds).

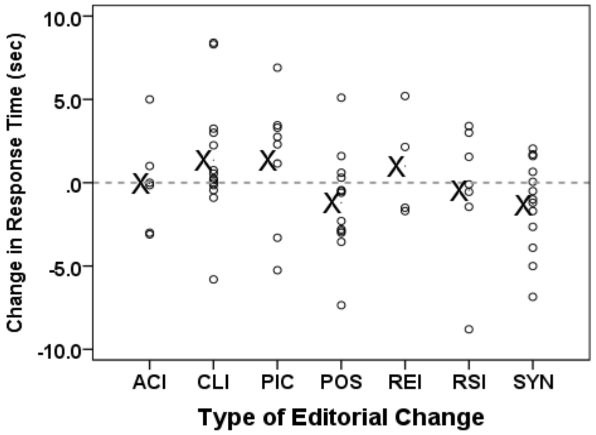

Figure 3. Change in median response time for different types of editorial changes.

ESL Examinees

All analyses were repeated for the subset examinees who indicated that they had learned English as a second language. In terms of item difficulty, none of the log-odds ratios for the seven classes of edits was statistically significant, nor were any of the *Q* tests. Odds ratios and CIs for the 65 individual item pairs were evaluated and two item pairs exhibited significant changes in difficulty. One item classified as *adding clarifying information* became easier with the revision (*p*-value increased from .64 to.78). The specific change was that BMI (body mass index) was added to the item stem; we could not identify a reason why adding BMI would impact item difficulty for this group on that item. Indeed, two other items to which BMI had been added became slightly more difficult (nonsignificant). The second item to show a significant change was in the *removal of possessives* category; that item inexplicably became more difficult (*p*-values decreased from.73 and .62). There were no significant changes in item discrimination (*Zr*) for ESL examinees for any of the classes of edits, and none of the *Q* tests reached statistical significance. In addition, none of the 65 individual item pairs had significant differences in *Zr*. However, there was a change in response time for ESL examinees. The *closed lead-in* class of edits just reached statistical significance, with a mean log RT of 0.067, indicating a slightly longer time to respond to the question format as opposed to incomplete statements. This is the same class of edits that produced a significant difference in RTs for the total group of examinees. None of the *Q* tests reached the level of statistical significance, and none of the 65 individual item pairs had significant differences in RTs for ESL examinees.

Discussion

Summary of Results

The following effects were observed at the level of class of edit or for the 65 individual item pairs within each class:

- As a whole, items in the *removal of possessives* category appeared to become slightly more difficult by dropping the apostrophe *s* from a diagnostic study or disease (mean difference in *p* = −0.016). However, none of the individual pairs of items exhibited a statistically significant difference in difficulty for the total group of examinees. One item in this class did become more difficult for ESL examinees. That edit involved changing *Meniere's disease* to *Meniere disease* in one of the distractors (original *p* = .734; revised *p* = .619).
- One item from the *closed lead-in* category became more difficult (original *p* = .934; revised *p* = .898). However, the items as a class did not exhibit a significant change in difficulty.
- There were no significant differences between original and revised items in terms of item discrimination (*r*, *Zr*) for either the total group or ESL examinees.
- As a class, the RTs for the *closed lead-in* category were significantly longer by 1.4 seconds for all examinees and 1.6 seconds for ESL examinees. The change in RTs for three of the individual item pairs reached statistical significance. The increase in RT for those three items ranged from 4.0 seconds to 8.9 seconds.
- There also were significantly longer RTs for one item involving *removal of possessives* (5.2 seconds longer) and for an item that added three words ("*in the diagram*") intended to direct examinees to a diagram was obviously displayed on the computer screen.
- A significant *Q* test suggested the presence of systematic variability in RTs for the class of edits involving the *removal of explanatory information*. The change in median response times for the four items in this class were −1.7, −1.5, 2.2 and 5.2 seconds. Three of the changes in this class were identical and involved dropping the parenthetical text from "*Pneumocystis jirovecii* (formerly *P. carinii*)."[1] The changes in median RTs for these items were −1.7, −1.5, and 2.2 seconds. The other change was to remove "(BUN)" from "urea nitrogen (BUN)," which posted a 5.2-second change in RT.
- There was no consistent evidence of differential effects for ESL examinees. The slightly longer reaction time for closed lead-ins applied to both native and nonnative speakers of English.

[1] This is an example of changing terminology. When such changes occur, both names are used, with the older term in parentheses, until such time as that the new term has become universally known.

Overall, the results indicate that the types of editorial changes made here had little or no systematic impact on item difficulty and perhaps a slight effect on response time. These findings are consistent with unpublished studies reporting that minor stylistic revisions have minimal impact on item performance (O'Neill, 1986; Webb & Heck, 1991; Zhang & Zhu, 2013). Although there was weak evidence for increased item difficulty for the *removal of possessives* category, there is no apparent explanation why this type of change would affect item difficulty. During the past 20 years there has been a trend in medical writing to remove possessives on eponyms (*AMA Manual of Style*, 10th Edition, 2007); however, both possessive and non-possessive forms are abundant in medical literature and well-known to examinees. The one item pair within the *closed lead-in* category that appeared to become slightly more difficult also defies explanation; it could reflect a real difference or might be Type I error.

One new and interesting finding was that the *closed lead-in* resulted in a slightly longer response time for all examinees – a plausible outcome given that *closed lead-ins* actually contain more words than open lead-ins, as illustrated in Table 1. Also, the distinguishing feature of the *closed lead-in* is the inclusion of a question mark, which may produce a more abrupt transition from stem to options than open lead-ins. While intriguing, this finding has limited practical application, given that RT typically is not factored into examinee scores and does not directly affect item difficulty or discrimination. One very important exception would be the circumstance in which numerous items were revised to the *closed lead-in* format on the same test form, which would presumably cause an increase in total test response time, which would then be expected to impact examinee performance on long and/or speeded tests.

Implications for Practice

The present findings contribute to a small but growing body of research indicating that items subjected to minor edits do not require re-pretesting. While the collective findings have immediate implications for test development, the practical challenge is that these studies have not exhaustively sampled the universe of possible edits. Thus, for those stylistic edits not studied, test developers must be able to accurately forecast whether a stylistic change will impact item performance. We informally tested this by asking three experienced editors to independently predict which editorial alterations would produce a change in item difficulty for the 65 item pairs. The editors were remarkably consistent and conservative in their judgments. All three editors flagged items that *added clarifying information* (ACI) and *removed explanatory information* (REI); one editor also flagged two of the eight items that *added text to items with graphics* (PIC). For the *adding clarifying information* and *removal of explanatory information* categories, a conservative approach to re-pretesting seems justified given that it is often difficult to determine a priori what constitutes a substantive change when adding or removing clarifyingor explanatory information. The results of the study and our three editors would indicate that re-pretesting should not be required for the other categories of stylistic edits. While the editors were more conservative than the data suggest is necessary, they were less conservative – and more accurate – than the conventional rule that all changes require re-pretesting.

Because the present study included a large sample of examinees, more item pairs, and more types of stylistic changes, the results encourage a more generalizable view than previous reports on the effects of minor editorial changes. No previous studies cited had explored the effect of the stylistic changes on response time, so our results in this area are particularly informative. Also, many previous studies included more extensive changes than those that we considered minor, such as correction of item flaws or addition or subtraction of clinical detail. The fact that several prior studies demonstrated that more substantive changes can affect item performance serves as an important reminder that there is some threshold above which changes do warrant re-pretesting.

Sample size is an issue for any study that fails to reject the null hypothesis. While the *N*s for the present study were only moderately large, combining results across items with similar types of edits increased statistical power and hopefully contributed to the generalizability of findings. However, larger sample sizes would have provided additional power to detect other possible differences that might exist. Furthermore, the classes of stylistic edits varied in terms of their internal similarity. While most classes are very homogeneous, others are not; *adding clarifying information* and *removal of explanatory information* are obviously heterogeneous and whether an edit makes a difference will depend on the specific information that was added or eliminated. The nonsignificant *Q* test generally supported aggregation, but the fact remains that the aggregated results may not have been particularly informative for these heterogeneous categories. This is one reason why we compared results at both the item level and the edit category level.

Additional research is warranted. This study included only those stylistic changes we felt were safe to label as minor edits. A new study, with input provided by subject matter experts, could include items for which a spectrum of minor to major changes is made. These study items would test content in well-established areas of medicine, thus eliminating other confounding factors, such as the emerging sciences, where the effect of examinee unfamiliarity with the content would be difficult to distinguish from the effect of the editorial changes. Where relevant, we would advocate that such studies be conducted with native and nonnative speakers of the particular language being studied. It also may be useful to investigate the effects of modifications prompted by new technologies, such as changes in screen sizes and displays or the introduction of hover text or zoom control capabilities. The cumulative findings of related research and the findings of the present study support a policy that does not require re-pretesting items that undergo minor stylistic changes. We would recommend that the informed judgments of subject matter experts and editorial staff be considered in deeming an edit major or minor within a systematic framework to ensure consistency. Clearly, there is a threshold above which changes warrant re-pretesting because prior studies demonstrated that more substantive changes can affect item performance; future research might seek to identify where that threshold lies.

References

Abedi, J (in press). Language issues in item development. In S. Lane, M.R. Raymond & T.M. Haladyna (Eds.), *Handbook of test development*, (2nd Ed). Mahwah, NJ: Lawrence Erlbaum Associates.

AMA Manual of Style: A Guide for Authors and Editors (10th Ed.). (2007). New York, NY: Oxford University Press.

Bornstein, H., & Chamberlain, K. (1970). An investigation of the effects of " verbal load" in achievement tests. *American Educational Research Journal*, 597-604.

Brennan, R. L. (1992). The context of context effects. *Applied Measurement in Education*, 5, 225-264.

Caldwell, D. J., & Pate, A. N. (2013). Effects of question formats on student and item performance. *American Journal of Pharmaceutical Education*, *77*(4).

Case, S. M., Swanson, D. B., & Becker, D. F. (1996). Verbosity, window dressing, and red herrings: do they make a better test item? *Academic Medicine*, *71*(10), S28.

Cassels, J.R.T., & Johnstone, A.H. (1984). The effect of language on student performance on multiple choice tests in chemistry. *Journal of Chemical Education* 61(7): 613.

Cizek, G. J. (1994). The effect of altering the position of options in a multiple-choice examination. *Educational and Psychological Measurement*, *54*(1), 8-20.

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education*, *10*(2), 133-143.

Dudycha, A. L., & Carpenter, J. B. (1973). Effects of item format on item discrimination and difficulty. *Journal of Applied Psychology*, *58*(1), 116.

Eva, K. W., Brooks, L. R., & Norman, G. R. (2001). Does "Shortness of Breath" = "Dyspnea"?: The Biasing Effect of Feature Instantiation in Medical Diagnosis. *Academic Medicine*, *76*(10), S11-S13.

Fleiss, J. L. (1994) Measures of effect size for categorical data. In H. Cooper and L.V. Hedges (eds.), *The handbook of research synthesis* (pp. 245-260). New York, NY: Russell Sage Foundation.

Green, K. (1984). Effects of item characteristics on multiple-choice item difficulty. *Educational and Psychological Measurement*, *44*(3), 551-561.

Haladyna, T.M., & Rodriguez, M.C. (2013). *Developing and validating test items*. New York: Routledge.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, linking, and scaling: Methods and practices*. New York, NY; Springer-Verlag.

Lipsey, M. W., & Wilson, D. (2000). *Practical meta-analysis*. Thousand Oaks, CA: Sage Publications.

Norman, G. R., Arfai, B., Gupta, A., Brooks, L. R., & Eva, K. W. (2003). The privileged status of prestigious terminology: impact of "medicalese" on clinical judgments. *Academic Medicine*, *78*(10), S82-S84.

O'Neill, K. A. (1986, April). *The Effect of Stylistic Changes on Item Performance.* Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA).

Plake, B. S., & Huntley, R. M. (1984). Can relevant grammatical cues result in invalid test items? *Educational and Psychological Measurement*, 44(3), 687-696.

Ratcliffe, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*, 510-532.

Schaap, L., Schmidt, H.G., & Verkoeijen, P.P. J. L. (2012). Assessing knowledge growth in a psychology curriculum: Which students improve most? *Assessment & Evaluation in Higher Education*, *37*(7), 875-887.

Tarrant, M., & Ware, J. (2008). Impact of item□ writing flaws in multiple□ choice questions on student achievement in high□ stakes nursing assessments. *Medical Education*, *42*(2), 198-206.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioural Statistics*, *31*, 181-204.

Violato, C., & Marini, A. E. (1989). Effects of stem orientation and completeness of multiple-choice items on item difficulty and discrimination. *Educational and Psychological Measurement*, *49*(1), 287-295.

Webb, L.C., & Heck, W.L. (1991, April). *The effect of stylistic editing on item performance.* Paper presented at the meeting of the National Council of Measurement in Education (Chicago, IL).

Young, M.J. (2006). Vertical scales. In S.M. Downing & T. M. Haladyna (Eds.). *Handbook of test development* (pp. 469-485). Mahwah, NJ: Lawrence Erlbaum Associates.

Zhang, Y., & Zhu, R. (2013, April). *The Impact of Minor Item Revision on Item Performance and Ability Estimation of an IRT-based Medical Certification Exam.* Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA).

Appendix: Analyses of Change in Item Difficulty, Discrimination, and Response Time

Table A.1: Item Difficulty by Class of Edit for All Examinees

Class of Edit		Mean *p* Value			Log-Odds Ratio		Odds Ratio		
	N	Original	Revised	Change	Mean	95% CI	Mean	95% CI	*Q*
ACI	6	.756	.748	−.008	−0.050	−0.170 to 0.070	0.951	0.843 to 1.072	5.02
CLI	14	.766	.757	−.009	−0.057	−0.141 to 0.027	0.945	0.869 to 1.028	10.58
PIC	8	.698	.708	.010	0.052	−0.051 to 0.155	1.054	0.950 to 1.168	0.90
POS	13	.780	.764	−.016	−0.091	−0.178 to −0.003*	0.913	0.837 to 0.997	4.17
REI	4	.779	.768	−.011	−0.061	−0.212 to 0.090	0.941	0.809 to 1.094	3.16
RSI	7	.784	.775	−.008	−0.053	−0.172 to 0.066	0.948	0.842 to 1.068	2.56
SYN	13	.782	.784	.003	0.021	−0.076 to 0.117	1.021	0.927 to 1.124	9.99

Note. Means and CIs for odds ratios obtained by back-transforming mean log-odds ratios. CI = confidence interval; ES = effect size; N = number of item pairs. * = statistically significant log-odds ratio or *Q* test.

Table A.2: Item-Total Correlation (r, Zr) by Class of Edit for All Examinees

Class of Edit		Mean Item-Total *r*			Change in Fisher's *Zr*		
	N	Original	Revised	Change	Mean ES	95% CI	*Q*
ACI	6	.238	.235	.003	.003	-.047 to .054	0.89
CLI	14	.266	.248	.018	.020	-.013 to .053	12.91
PIC	8	.236	.240	-.004	-.004	-.050 to .041	1.07
POS	13	.283	.288	-.005	-.005	-.041 to .030	12.63
REI	4	.228	.260	-.032	-.034	-.096 to .028	2.75
RSI	7	.224	.240	-.017	-.018	-.067 to .030	4.44
SYN	13	.233	.227	.005	.008	-.028 to .043	16.63

Note. CI = confidence interval; ES = effect size; N = number of item pairs. * = statistically significant Fisher's *Zr* or *Q* test (none significant).

Table A.3: Response Time by Class of Edit for All Examinees

Class of Edit		Response Time (in sec)			Change in Log Response Time		
	N	Original	Revised	Change	Mean ES	95% CI	*Q*
ACI	6	71.5	71.5	0.0	.010	−.040 to .061	10.32
CLI	14	61.0	62.4	1.4	.037	.004 to .070*	19.62
PIC	8	56.1	57.6	1.5	.019	−.027 to .064	11.28
POS	13	59.4	58.1	−1.3	−.011	−.047 to .025	17.84
REI	4	59.9	60.9	1.0	.031	−.031 to .093	7.98*
RSI	7	74.1	73.7	−0.4	.013	−.035 to .061	7.50
SYN	13	67.6	66.3	−1.3	.012	−.023 to .047	3.55

Note. CI = confidence interval; ES = effect size; N = number of item pairs. * = statistically significant log response time or *Q* test.

Citation:

Stoffel, Heather, Raymond, Mark R., Bucak, S. Deniz & Haist, Steven A. (2014). Editorial Changes and Item Performance: Implications for Calibration and Pretesting *Practical Assessment, Research & Evaluation*, 19(14). Available online: http://pareonline.net/getvn.asp?v=19&n=14

Corresponding Author:

Mark Raymond
Director, Research and Development
Test Development Services
National Board of Medical Examiners
Mraymond [at] nbme.org