Editorial Changes and Item Performance: Implications for Recalibration and Pretesting

Heather Stoffel, Mark Raymond, S. Deniz Bucak, Steve Haist

National Board of Medical Examiners

Paper presented at the 2014 annual meeting of the

American Educational Research Association (AERA)

Contact: Heather Stoffel

hstoffel@nbme.org

Editorial Changes and Item Performance: Implications for Recalibration and Pretesting

Abstract

Previous research on the impact of text and formatting changes on test-item performance has produced mixed results. The present study investigated the effects of 7 classes of stylistic changes on item difficulty and response time for native English speakers and English as a Second Language speakers. Each of 65 item pairs (original and revised) were assigned to approximately 480 examinees per item in a randomized experiment. None of the changes affected item difficulty, while 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. These results have implications for the conventional practice of re-pretesting (or recalibrating) items that have been subjected to editorial changes.

Editorial Changes and Item Performance: Implications for Recalibration and Pretesting

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 (Kolen & Brennan, 2004). 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. Many psychometricians also advocate that a revised item – whether or not it explicitly functions as a linking item – be treated as a new item. Testing agencies often apply this rule to any type of revision, from minor edits, such as replacing commas with semicolons, to more extensive changes, such as reordering options (Cizek, 1994; Kolen & Brennan, 2004).

The present study was undertaken in part to evaluate the "revised item means new item" rule. While application of that rule is certainly safe, it may not always be prudent; re-pretesting is costly because it displaces the capacity to pretest original material and adds to the time required for an item to become operational for scoring purposes. This disadvantage has become compounded in recent years because test materials now require more frequent revision than ever before. In some instances, revisions are made because of a change in authoritative style guidelines (e.g., *The Chicago Manual of Style*; *American Medical Association Manual of Style*). In other cases, the evolving language of medical practice induces the changes. It is now common, for example, to refer to "human immunodeficiency virus infection" as simply "HIV infection" in medical text, as knowledge of and reference to the disease increased over time. Changes in classification schemes (such as the renaming of microorganisms) stimulate revision, as does changing technology. For example, the replacement of analog medical imaging with digital imaging required that many credentialing agencies remove the word *film* from test questions about x-rays. In this case, the previous term and the new term are understandable without additional explanation or clarification, as are many other minor style or editorial changes. The decision for test developers is whether to continue with the old style or update hundreds or even thousands of test

questions to reflect the new style. Assuming the change is imperative, the next step is to decide whether re-pretesting is really necessary.

Consistent application of the "revised item means new item" rule seems judicious because test items often fail to perform as predicted, and even minor edits have been shown to impact item statistics. Indeed, research on context effects has identified instances in which an item's performance is affected by its location relative to other items (Brennan, 1992). The factors that might affect item difficulty are so numerous that it is often challenging to distinguish the cause of performance changes. On the one hand, it seems obvious that extensive changes to the wording or structure of an item will affect item performance. Examples of such changes might include adding clarifying information to explain complex terminology, adding graphics, reordering option sets, replacing a distractor, or revising items to eliminate item-writing flaws. 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.

Studies have examined the impact of adding or removing medical information from items. One investigation showed that adding or taking away information (e.g., patient history, diagnostic studies) in a vignette does have an effect on item difficulty and discrimination (Case, Swanson, & Becker, 1996). These results are echoed by studies comparing examinee performance on examinations containing items in which medical terminology and lay terminology were interchanged in various permutations (Eva, Brooks, & Norman, 2001; Eva & Wood, 2003; Eva, Wood, Riddle, Touchie, & Bordage, 2010; Norman, Arfai, Gupta, Brooks, & Eva, 2003); results showed that these types of edits did have an effect on performance but in different ways, based on examinee characteristics such as inherent candidate ability and international medical graduate status. A study of the effect of changes in option set order (Cizek, 1994) showed significant but unpredictable changes in item difficulty. Reshetar, Mills, Norcini, and Guille (1998) found that altering key details of patient vignettes and substituting options had a variety of effects on *p* and *r* (correlation coefficient) values. The types of changes made in all of these studies are considerable and certainly would be expected to impact item performance. In a similar vein, studies of alterations in language difficulty have demonstrated an effect on item statistics (Cassels & Johnstone,

1984; Plake & Huntley, 1984); however, others have shown only small effects (Bornstein & Chamberlain, 1970; Green, 1984). Grammatical changes and alterations in language difficulty appear to be complicated and the results suggest that re-pretesting is advised.

In contrast, a few reports, all of them unpublished, have suggested that many types of revisions have little or no impact on the statistical characteristics of test items. O'Neill (1986) found no significant differences in performance on item pairs for which small changes in abbreviations, symbols, or drug names (i.e., generic vs. proprietary) were made. A later study by Webb & Heck (1991) offered further support that style changes had no detectable effect on item difficulty. Most recently, Zhang & 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 showed that these types of changes had little impact on item performance. These results support the intuition that re-pretesting is not required with certain types of minor edits.

Because items are often revised to correct specific types of item-writing flaws, there has been interest in studying the impact of these types of edits on item performance. Eliminating agreed-upon item-writing flaws (Haladyna, 2004), such as use of none-of-the-above, negatively phrased lead-ins, and option convergence, has been shown to affect item difficulty and overall examinee performance (Caldwell & Pate, 2013; Cassels & Johnstone, 1984; Downing, 2005; Dudycha & Carpenter, 1973; Green, 1984; Tarrant & Ware, 2008). Similarly, item-writing guidelines advocate the direct-question format over an incomplete stem for which each option completes the lead-in sentence (Haladyna, 2004). Studies comparing the two formats have reported an effect on item difficulty (Bolden & Stoddard, 1980; Violato & Marini, 1989).

Useful comparison of many of these studies is problematic due to radically different examinee populations (ranging from elementary and high school students to medical school candidates) and small sample sizes. It seems clear from the accumulated evidence that the degree to which a textual change affects cognitive processing does matter in deeming an edit minor or major, as do grammar changes and the removal of item flaws. Notably, studies suggest that the impact of changes to text also depends on

language fluency, with examinees from foreign countries doing better with more formal terminology (Eva et al., 2010). Meanwhile, other studies suggest that truly minor changes have minimal impact on item statistics (O'Neill, 1986; Webb & Heck, 1991, Zhang & Zhu, 2013). However, the findings are equivocal and sometimes unpredictable.

The purpose of this research was to experimentally determine the extent to which different types of minor editorial changes affect item performance. This study built on previous research in four ways. First, we included a large sample of items (65 pairs) representing various types of editorial changes. 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, both item difficulty and time required to respond to each item were studied. Both dependent measures are useful because it is possible that certain editorial or stylistic changes could affect reading time without affecting item difficulty. Finally and most importantly, we studied the impact of changes on the performance of a subset of examinees who were not native speakers of English.

Method

Data Sources

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 for this study and did not count towards the examinees' scores. Each item pair consisted of an original and a revised version. These 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 could be closer to the proposed line between major and minor revisions.

Pretest items were distributed across test forms and examinees such that each examinee saw a stratified 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 study 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. Each item pair can be regarded as a replication across independent samples, with each replication being on a different scale determined by the content and difficulty of that item pair.

Analyses

Descriptive statistics and inferential tests are reported for both item difficulty and response time (RT). Significance testing was done at two levels. First, data were aggregated within each of the seven edit classes using meta-analytic methods (Hedges & Olkin, 1985; Lipsey & Wilson, 2000). Mean effect sizes (i.e., change in difficulty 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. Second, CIs were computed for each of the 65 item pairs to determine if there was a change in item difficulty or response time. CIs that did not include zero were regarded as statistically significant.

Item difficulty. First, item means (*p* values) were obtained and the difference in *p* values was plotted for each item pair. Next, odds ratios were computed and served as the basis 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. To facilitate interpretation, log-odds ratios also were transformed back to the odds ratio scale. After evaluating odds ratios within each class of edits, odds ratios for the 65 pairs were inspected.

Response time (RT). The time, in seconds, for each examinee to respond to an item was recorded. For descriptive purposes, we report median response time 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 a large number of statistical tests were done, the exploratory nature of this study and our willingness to tolerate declaring a change as significant when it was not (i.e., false-positives) warranted a liberal approach to significance testing.

Results

Item Difficulty

All examinees. Figure 1 shows the change in *p* value for each item within the seven classes of edits, with positive values indicating that the revised item was easier (i.e., had a higher *p* value) than the original item. The changes in *p* values for individual items range from about −0.05 to 0.04, with standard errors of the differences ranging from 0.008 to 0.033 and averaging about 0.025. In Figure 1, within each column, the X's correspond to the mean for each class of edit. The largest within-class mean difference is for *removal of possessives* (POS) (i.e., dropping an apostrophe *s*), with a mean change of −0.016.

The log-odds ratio for each item pair was weighted by the inverse of its standard error for cumulating effects across replications (Hedges & Olkin, 1985), and the mean effect size and 95% CIs

were obtained within each of the seven classes of edits. Table 2 shows the results, along with the original *p* values. To facilitate interpretation, the mean log-odds ratios and CIs were transformed back into odds ratios. The *Q* indices, which evaluate the homogeneity of log-odds ratios, are also presented.

The only type of edit for which the log-odds ratio failed to include zero in the 95% CI was *removal of possessives*, which barely reached statistical significance. The CIs for six remaining classes of edits all contained zero, indicating that the difficulty of the item pairs is not significantly different. None of the *Q* tests reached statistical significance; this indicates that the hypothesis of homogeneity of log-odds ratios cannot be rejected within any of the seven classes of edits and further suggests that any variation in changes in item difficulty is due to sampling error. For completeness, odds ratios for individual item pairs also 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 odds ratios within the *removal of possessives* category was significant.

ESL examinees. The analyses were repeated for examinees who indicated that they had learned English as a second language; Figure 2 presents the change in *p* values for this group. The values for the 65 individual pairs are distributed in a fairly symmetric manner around the value of 0, as are the means for each edit category. Compared with Figure 1, there is greater variability in *p*-value differences, which can be attributed in part to the smaller sample sizes. The average number of ESL examinees responding to each item was 154, and standard errors for differences in *p* values ranged from 0.015 to 0.061, with an average of 0.047.

Table 3 presents the aggregated effect sizes and related statistics within each class of edits for ESL examinees. None of the odds ratios reached statistical significance, although *closed lead-in* and *removal of possessives* were at the borderline ($0.05 < p < .10$). None of the *Q* tests was statistically significant, indicating that the variation in changes in item difficulty is not greater than what would be expected due to sampling error. Odds ratios and CIs for the 65 item pairs were evaluated and an item classified in the *removal of possessives* category was found to be significant.

Response Time

All examinees. Analyses of RTs mirrored those for item difficulty except that medians were used in graphic summaries, while log RTs served as the basis for computing and cumulating effect sizes. Figure 3 shows the change in median RT, in seconds, for the 65 item pairs. The differences ranged from an 8.4-second increase to a 10.4-second decrease, with most changes (54 of 65) falling within ± 5 seconds. Figure 3 does suggest that RTs are slightly longer for *closed lead-in* and possibly for *removal of explanatory information* (REI). However, the average increase in response time for these two categories is only about 1.4 seconds.

To more formally evaluate these differences, log RTs were converted to effect sizes and combined across all items within each class of edit. The results are summarized in Table 4. The CI for *closed lead-in* does not include zero, indicating that its effect size of 0.037 was significantly different from zero. The median difference in response time for *closed lead-in* was 1.4 seconds, with a difference in log RT of 0.023, suggesting that direct questions (revised version) required slightly longer response time compared with incomplete statements or open-ended lead-ins (original version). Table 4 also indicates that the 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.

Confidence intervals corresponding to the mean change in log RTs for the 65 individual item pairs were computed; the CIs for six item pairs did not include 0 and were flagged as statistically significant. 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). 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).

ESL examinees. Figure 4 presents the differences in median RTs for ESL examinees. As expected, due to smaller sample sizes, there is greater variability in the 65 individual means for ESL examinees when compared with Figure 3.

Table 5 presents the mean effect sizes and CIs for each class of edit. It can be seen that the mean changes in RT are quite small as are the mean effect sizes associated with each. There is, however, one difference that just reached statistical significance: the 95% CI for *closed lead-in* does not include zero, indicating a longer time was required to respond. It should be noted that this same class of edits produced a significant difference in RTs for the total group of examinees. None of the *Q* tests reached the level of statistical significance required ($p < 0.05$) to reject the hypothesis for homogeneity of effect sizes, and none of the 65 individual item pairs had significant differences in RTs for ESL examinees.

Discussion

Summary of Results

The following effects were seen across the 65 item pairs:

- As a class, items in the *removal of possessives* category became slightly more difficult by dropping the apostrophe *s* from a diagnostic study or disease (mean difference in $p = -0.016$). However, none of the pairs of items exhibited a statistically significant difference in difficulty for the total group of examinees. Across all 13 items, the largest (but nonsignificant) difference in *p* value for all examinees was for an item that had been changed by dropping the apostrophe *s* from *Gram's stain* in a distractor. One item became 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 $p = .734$; revised $p = .619$). However, the items as a class did not exhibit a significant change in difficulty.
- As a class, the RTs for the *closed lead-in* category were significantly slower by 1.4 seconds for all examinees and 1.6 seconds for ESL examinees. This difference (1.4 vs 1.6 seconds) cannot be regarded as significant. 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 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),"[2] 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.

Overall, the results indicate that the types of editorial changes made in this study had little or no systematic impact on item difficulty and perhaps a slight effect on response time. These findings are consistent with previous (unpublished) research papers reporting that minor stylistic changes 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 logical reason 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 a few more words than open lead-ins, as illustrated in Figure 5. 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 does not directly affect examinee scores or the types of item parameter estimates typically used for scoring and equating. 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 could cause an increase in total test response time, which would then 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 edit types. Thus, for those stylistic edits not studied, test developers must be able to accurately forecast whether a stylistic change will impact item performance. For the 65 items used in the present study, we asked three experienced editors to independently predict which editorial alterations would produce a change in item difficulty. The editors were remarkably consistent and conservative in their judgments, with each flagging from nine to 11 items spanning three classes of edits: *adding clarifying information* (five of six item changes flagged) and *removal of explanatory information* (four of four flagged), and *adding text to items with graphics* (PIC) (two of eight flagged by one editor). 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 clarifying/explanatory information. Results for *closed lead-in* showed only a small influence on response time and no effect on item difficulty, and none of the editors thought that this would result in a change in difficulty; therefore, re-pretesting is not warranted. Although results showed weak evidence for increased item difficulty for the *removal of possessives* category, we suspect that the significant results are spurious and that these changes do not require re-pretesting. Both the data and our three editors generally concurred that re-pretesting should not be required for the other categories of stylistic edits (*adding text to items with graphics*; *replacing term with synonym* (SYN)*;* and *removal of superfluous information* (RSI)). In summary, 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 a reminder that there is some threshold above which changes do warrant re-pretesting.

The sample sizes for the present study, while reasonable, were not large. While combining results across items with similar types of edits increased statistical power and contributed to the generalizability of findings, 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 were 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 homogeneous 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, thus eliminating the need for re- pretesting. 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 compounding 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 may also 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

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

Bolden, B. J., & Stoddard, A. (1980, April). *The Effects of Language on Test Performance of Elementary School Children.* Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA).

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.

Eva, K. W., & Wood, T. J. (2003). Can the strength of candidates be discriminated based on ability to circumvent the biasing effect of prose? Implications for evaluation and education. *Academic Medicine*, *78*(10), S78.

Eva, K. W., Wood, T. J., Riddle, J., Touchie, C., & Bordage, G. (2010). How clinical features are presented matters to weaker diagnosticians. *Medical Education*, *44*(8), 775-785.

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. (2004). Developing and validating multiple-choice test questions. Mahwah, NJ: Lawrence Erlbaum Associates.

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.

Reshetar, R., Mills. L., Norcini. J., & Guille, R. (1998). Is it worth the time to attempt fixing an item? In *Ottawa Conference in Medical Education.* (Philadelphia, PA).

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).

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).

Footnotes

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

[2]This is an example of a change made in the interest of language precision. While "urea nitrogen (BUN)" is widely used, it is not accurate because the concentration of this substance is measured serum, not in blood.

Table 1

Classes of Editorial Changes

Code	N	Class of Edit	Explanation and/or Examples
ACI	6	Adding clarifying information	Include additional information, sometimes in parentheses. For example, adding BMI to existing height and weight information.
CLI	14	Closed lead-in	Change stem from open ended (each option completes the stem) to the interrogative form ending with a question mark.
PIC	8	Adding text to items with graphics	Rather than just displaying a graphic, change text to explicitly say "in the photograph shown."
POS	13	Removal of possessives	Remove apostrophes from eponyms. For example, "Wilson disease" instead of "Wilson's disease."
REI	4	Removal of explanatory information	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.
RSI	7	Removal of superfluous information	Remove information that has become obsolete, such as "film" from "x-ray film."
SYN	13	Replacing term with synonym	Interchange essentially synonymous terms, such as "limbs" with "extremities" or "neonate" with "newborn."

Note. N = number of item pairs.

Table 2

Item Difficulty by Class of Edit (CE) for All Examinees

		Mean *p* Value			Odds Ratio		Log-Odds Ratio		
CE	N	Orig	Revised	Change	Mean ES	95% CI	Mean ES	95% CI	*Q*
ACI	6	.756	.748	−.008	.951	.843 to 1.072	−.050	−.170 to .070	5.02
CLI	14	.766	.757	−.009	.945	.869 to 1.028	−.057	−.141 to .027	10.58
PIC	8	.698	.708	.010	1.054	.950 to 1.168	.052	−.051 to .155	.90
POS	13	.780	.764	−.016	.913	.837 to .997	−.091	−.178 to −.003*	4.17
REI	4	.779	.768	−.011	.941	.809 to 1.094	−.061	−.212 to .090	3.16
RSI	7	.784	.775	−.008	.948	.842 to 1.068	−.053	−.172 to .066	2.56
SYN	13	.782	.784	.003	1.021	.927 to 1.124	.021	−.076 to .117	9.99

Note. CI = confidence interval; ES = effect size; N = number of item pairs. * = statistically significant.

Table 3

Item Difficulty by Class of Edit (CE) for English as a Second Language Examinees

		Mean *p* Value			Odds Ratio		Log-Odds Ratio		
CE	N	Orig	Revised	Change	Mean ES	95% CI	Mean ES	95% CI	*Q*
ACI	6	.721	.730	.010	1.040	.845 to 1.280	.039	−.168 to .247	9.77
CLI	14	.704	.693	−.011	.945	.822 to 1.086	−.057	−.196 to .082	12.64
PIC	8	.665	.668	.004	1.018	.852 to 1.216	.018	−.160 to .195	5.68
POS	13	.755	.733	−.023	.888	.764 to 1.033	−.119	−.270 to .032	6.87
REI	4	.734	.756	.022	1.126	.871 to 1.455	.119	−.138 to .375	.91
RSI	7	.761	.754	−.007	.968	.789 to 1.187	−.033	−.237 to .171	2.37
SYN	13	.751	.746	−.005	.965	.819 to 1.137	−.036	−.200 to .128	13.58

Note. CI = confidence interval; ES = effect size; N = number of item pairs.

Table 4

Response Time by Class of Edit (CE) for All Examinees

		Average Response Time			Log Response Time		
CE	N	Orig	Revised	Change	Mean ES	95% CI	*Q*
ACI	6	71.5	71.5	.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	1.4	.019	−.027 to .064	11.28
POS	13	59.4	58.1	−1.2	−.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	−.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.

Table 5

Response Time by Class of Edit (CE) for English as a Second Language Examinees

		Average Response Time			Log Response Time		
CE	N	Orig	Revised	Change	Mean ES	95% CI	*Q*
ACI	6	73.1	73.9	.8	−.006	−.096 to .084	8.08
CLI	14	63.4	64.9	1.6	.067	.008 to .125*	6.09
PIC	8	58.0	58.4	.4	−.033	−.113 to .047	2.29
POS	13	61.0	60.5	−.5	−.015	−.079 to .048	6.89
REI	4	62.4	62.3	−.1	−.015	−.125 to .095	.43
RSI	7	76.3	76.1	−.3	.001	−.086 to .087	1.14
SYN	13	70.0	69.2	−.8	−.007	−.070 to .055	5.84

Note. CI = confidence interval; ES = effect size; N = number of item pairs.

* = statistically significant.

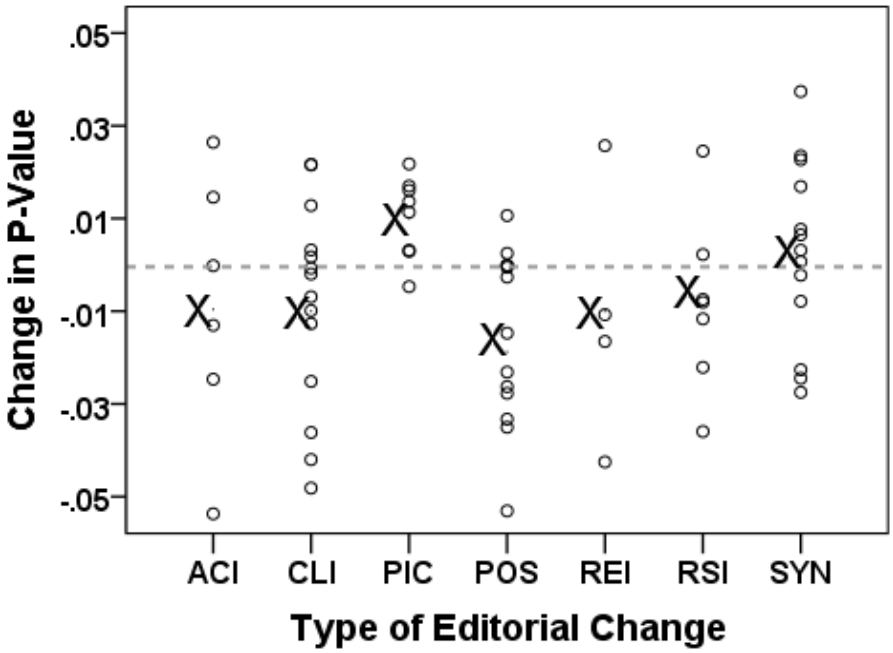


Figure 1. Change in *p* values for different types of editorial changes – all examinees.

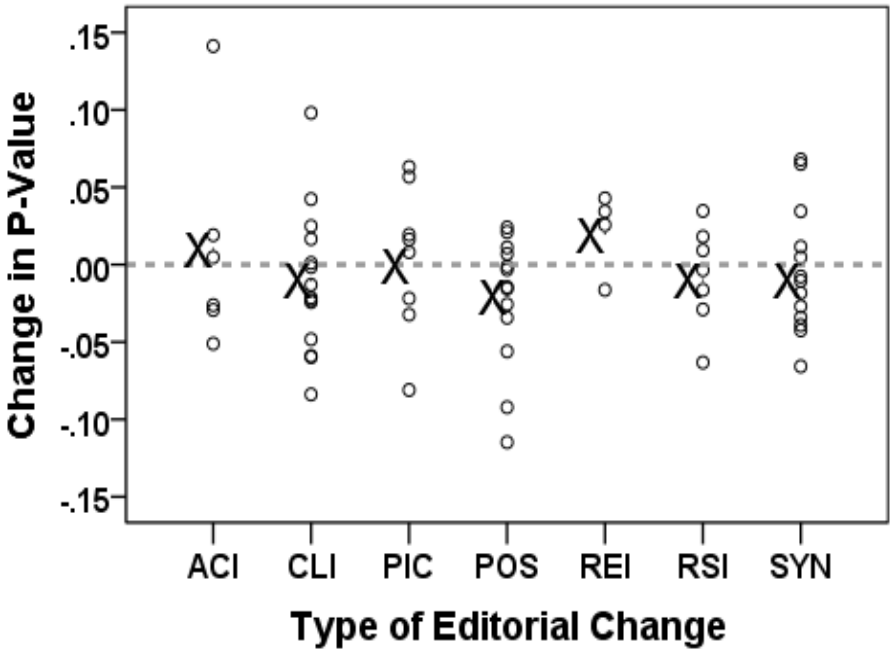


Figure 2. Change in *p* values for different types of editorial changes – English as a second language examinees.

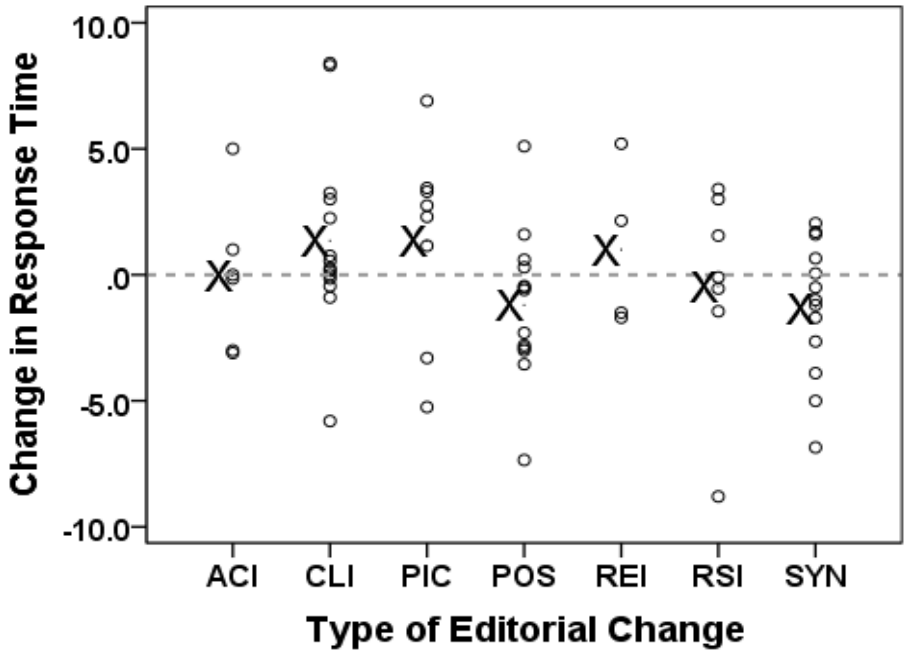


Figure 3. Change in median response time for different types of editorial changes – all examinees.

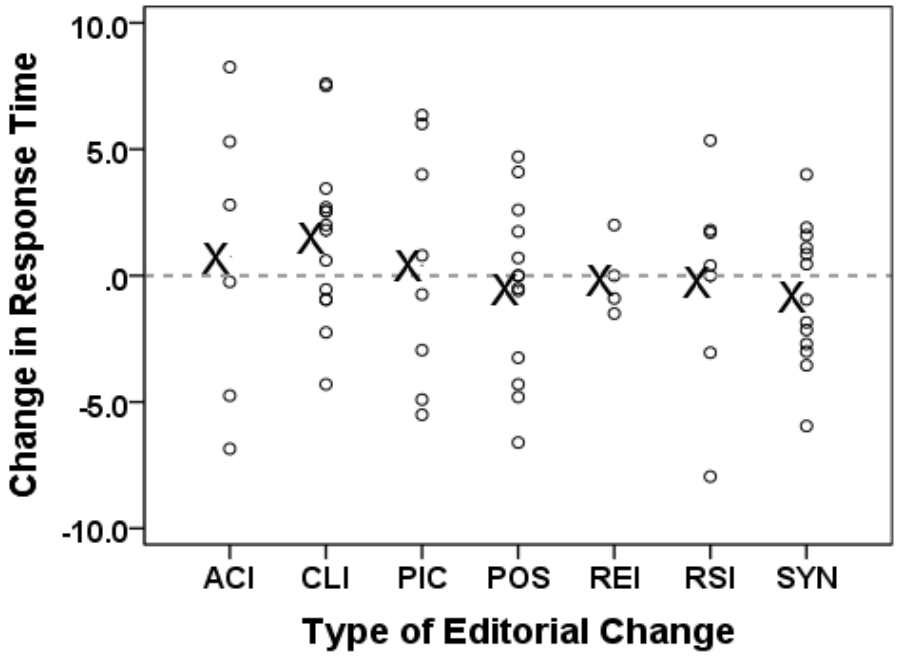
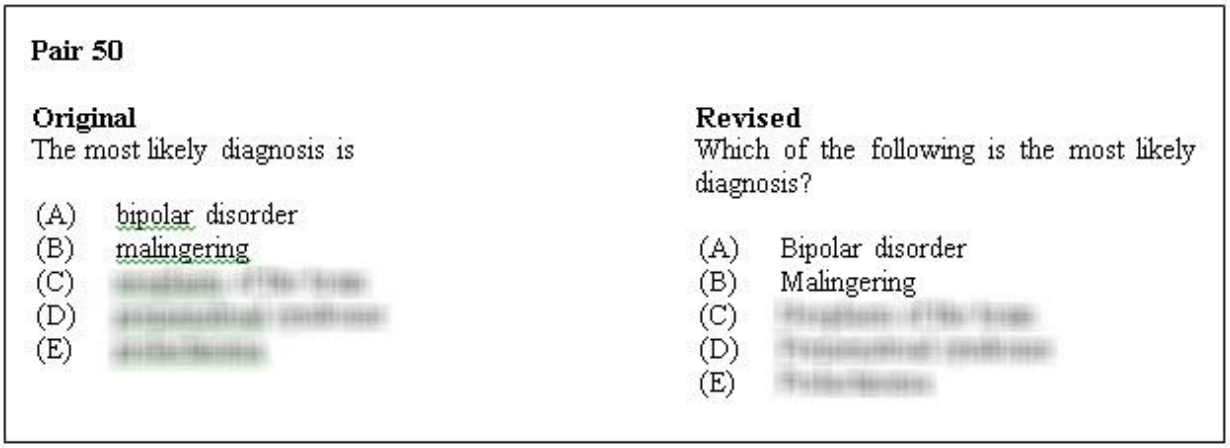


Figure 4. Change in median response time for different types of editorial changes – English as a second language examinees.



Pair 50

Original
The most likely diagnosis is

(A) bipolar disorder
(B) malingering
(C) [illegible]
(D) [illegible]
(E) [illegible]

Revised
Which of the following is the most likely diagnosis?

(A) Bipolar disorder
(B) Malingering
(C) [illegible]
(D) [illegible]
(E) [illegible]

Figure 5. An example of a closed lead-in (CLE) type of edit.