Reply to Commentary

Colby replies to Frederick

In replying to Frederick's remarks, I will start by thanking him for taking the time to respond with his thoughts and perspectives. One does not have to look far to know that Frederick is one of the more prolific and influential writers about validity assessment in the neuropsychological and forensic areas.

Perhaps the best place for me to begin is where he ended. Insightfully, Frederick identified a problem with all assessment instruments like the TOMM [1], namely, that it/they may be more tests of intention than of effort. An even more accurate label might be "confidence measure", since the primary purpose is to provide information about how much confidence may be placed in the reliability of the rest of the test results.

Frederick's contrasting effort and intention shows a sensitivity to the ways other areas of psychology might helpfully inform the science of clinical assessment. One can easily envisage a series of well-controlled, laboratory studies contrasting the effects of these two hypothetical constructs upon some dependent variables of interest. It is certainly a worthy challenge to the scientific community.

In response to Frederick's suspicion that most people who do poorly on the TOMM intend to do so, the clinical patient data reported in the test manual argue otherwise. Assuming that Frederick meant "more than half" when he said "most", can we assume that he believes that more than half of the clinical patients reported in the test manual failed the TOMM by intent? If not, then the bases for his suspicion are obscure. The real problem with tests like the TOMM is not how the members of norming groups perform; it is how persons perform about whom the examiner has little information. Using the normative data from the test manual as bases for comparison, my paper offered some different ways of thinking about the TOMM (and, by corollary, other similar assessment instruments) which might be useful for examiners faced with ambiguous testing situations.

In response to Frederick's first criticism, I did not state that the probabilities of correct and incorrect answers in a two-stimulus forced choice test, given equally tenable choices due to the absence of ability to discern them, were unequal, although I have previously commented upon why this hypothesis likely should never have been used, in the first place, for these types of tests [2]. What I stated here is that since cognitively impaired persons repeatedly have scored better then 50% correct on these types of tests, then it should be obvious that, for these tests, the probabilities of correct and incorrect answers were a priori not equal. For unknown reasons, Frederick did not comment upon my discussion of what "pure chance" might mean in the particular situation of impaired persons giving their best efforts on tests like this, choosing, instead, only to refer to a fair coin example.

To use a different metaphor than coin tosses, if a population's long-run average (i.e., expected) ability is 95 out of 100 correct (i.e., $p = 0.95$), then guessing the performance on any given item is similar to drawing a ball from a bag containing 95 black balls and 5 white balls, replacing the drawn ball after each draw (i.e., sampling with replacement). Each draw is still "pure chance", but the a priori probabilities are $p = 0.95$ for a black ball and $q = 0.05$ for a white ball on each draw. Frederick is correct in stating that using fair coin probabilities to gauge effort (his term was malingering) is an unfruitful venture; my point was that for researchers and test publishers even to have assumed that it might be fruitful, and then to discard use of the binomial distribution, in general, as unfruitful, is what is most unfortunate.

Coming to the TOMM's defense, Frederick stated that focusing upon "patterns of responding" have proved to be more fruitful for identifying how intact, impaired, and feigning patients respond than focusing upon p and q. Although I am unaware of any published research on inter-item response patterns on the TOMM, I definitely agree that the test is useful in discriminating between true and feigned impairment. I merely suggest that different decision rules should be used to evaluate test performance in order to increase specificity without unduly sacrificing sensitivity.

ISSN 1053-8135 / $8.00 © 2001, IOS Press. All rights reserved

Perhaps readers will decide that Frederick and I are quibbling over terminology in differing about what the proper null and alternate hypotheses should be in assessing effort (or intention). However, choosing a correct null hypothesis, if one takes a scientific approach to the assessment of effort (or intention), is more than just a "clever twist", to use Frederick's phrase. It is critical. A null hypothesis states, by definition, that there is no difference between a test individual's score and that of the average comparison person. For the TOMM and tests like it, the correct comparison person is not, and never should have been, an hypothetical person with "absence of any ability". It is the impaired patient who intends to do well and tries hard to do so. From his comment, Frederick apparently refers to this type of comparison as using a "floor effect" decision strategy.

Although it may sometimes be accurate to call the TOMM and similar tools "floor effect" instruments, depending upon the characteristics of the comparison populations used, choosing the correct null hypothesis requires using the correct comparison population. For any psychological test where there is but one correct answer for each item, the distribution of scores follows a binomial rule. If the probabilities of correct responses on individual items differ, the shapes of the binomial distributions may be more bell-shaped than if they are equal, as is generally assumed to be true for the TOMM.

Given how strongly he critiqued other aspects of my article, to my surprise, Frederick may have missed its main point, namely, that using frequency distributions like the binomial to derive cut scores which correspond to a priori specificities is a much more scientific way of making decisions in test situations than using cut scores which have been derived from the performances of arbitrarily gathered convenience samples. This is true for tests like the TOMM as well as for tests of actual ability. It was by using these techniques that I developed some alternate decision rules for the TOMM, by way of example, which would minimize both false positive and false negative errors.

As is true of any situation where comparisons to published norms are made, using the kind of decision rules I have proposed requires making some decisions about what comparison populations to use for particular testing situations. It is my position that such decisions are better made with an underlying scientific methodology in mind than simply with automatic acceptance of prescribed decision rules which have been shown to produce unacceptably high false positive rates among some well-defined clinical populations. Whether making and then testing these assertions empirically, using published data for bases of comparison, was unnecessary and muddled will be for individual readers to decide.

Faulder Colby, PhD
Oregon Health Sciences University

References

[1] T.N. Tombaugh, *Test of Memory Malingering: TOMM*, Multi-Health Systems, Inc., North Tonawanda, NY, 1996.

[2] F. Colby, Does the binomial distribution stand falsely accused? *Brain Injury Source* **4** (2000), 18–21.