

Review of the Inventory of Legal Knowledge¹

Author: Steve Rubenzer, Ph.D., ABPP, MHM Services, Office of the Forensic Examiner, Concord, NH. Email: rubenzer.steve@att.net

Abstract

The Inventory of Legal Knowledge (ILK) is a new instrument designed to assess feigning in the context of a competency-to-stand-trial evaluation. The test's development, rationale, user-friendliness, and implementation are discussed and critically reviewed. Several shortcomings in the validity evidence are revealed, suggesting the ILK's ability to discriminate malingerers from those actually incompetent may be much less impressive than implied in the manual. The recommended cutoff score of 47 produces unacceptable levels of false positives among defendants classified as actually incompetent and should not be used.

Keywords: Inventory of Legal Knowledge, validity testing, malingering, review

The *Inventory of Legal Knowledge* (ILK; Musick & Otto, 2010; Otto, Musick, & Sherrod, 2010) is a 61-item interview measure of response style. Unlike nearly all other such validity measures, it focuses on knowledge of the criminal justice system. Although its items form a face-valid, true-false test of legal knowledge, the authors state the ILK is not to be used in this way. Rather, it is used to determine whether a defendant is feigning lack of legal knowledge, presumably as a strategy to malingering incompetence to stand trial.

I initially questioned the value of such an approach. Having performed nearly 3,500 competency evaluations in Texas, I found it quite rare that a defendant would confine malingering to the domain of legal knowledge. More typically, defendants would feign just about any impairment they could think of: psychosis, mental retardation, *and* ignorance of the court system. I felt that assessment of malingering could easily be handled by standard tools for this purpose such as the Miller Forensic Assessment of Symptoms Test (M-FAST; Miller, 2001), MMPI-2-RF (Ben-Porath & Tellegan, 2008), Structured Inventory of Reported Symptoms-2 (Rogers, Sewell, & Gillard, 2010), Word Memory Test (Green, 2005), and Test of Memory Malingering (TOMM; Tombaugh, 1996). How-

¹ Editor's Note: Thanks to Randy K. Otto and Jeffrey E. Musick for checking an earlier draft for factual errors.

ever, I recently took a position in New Hampshire with the Office of the Forensic Examiner, and found a substantially different population and style of presentation. Among the other differences, and there were a significant portion of people who presented as having a serious deficit in their knowledge of the court system without presenting as grossly impaired in other ways. One even passed multiple symptom validity tests. It became apparent that feigning lack of knowledge was a popular strategy for malingering incompetence in my new locale, more so than claiming to hear voices.

Format and Suitable Populations

The ILK is administered as a structured interview. There is a test booklet that contains the test questions, and a separate, carbonless form for recording answers. For each question, the examinee must answer “true” or “false” and the examiner circles the appropriate answer on the scoring sheet. Scoring is a simple matter of tallying correct responses.

ILK scores can be interpreted by two approaches. Significantly below-chance performance has long been considered the gold standard in psychometric malingering assessment (Berry & Schipper, 2008; Boone, 2007; Slick, Sherman, & Iverson, 1999; Sweet, Condit, & Nelson, 2008), and ILK scores below 24 exceed the 95% confidence interval for chance-level performance. Perhaps as important, data presented in the ILK manual suggest that nearly 50% of actual malingerers obtain scores this low—a much higher percentage than reported for many other instruments relying on this strategy. Scores on the ILK can also be interpreted by reference to several normative groups.

The ILK is designed for English-speaking defendants at least 12 years of age who can comprehend spoken language at a 5th grade level. The authors report the ILK has not been specifically validated for individuals with significant cognitive limitations, and that caution should be used in interpreting results for such individuals. However, below-chance performance, when it occurs, remains very solid evidence of feigning regardless of intellectual level.

Construction

The ILK items were created by the authors to represent a broad range of content relevant to competency to stand trial. More than one hundred potential items were created and grouped into eleven content categories. The authors then culled items that were judged excessively trivial or technical, confusing, or redundant. Item language was refined to increase clarity and items were assessed using the Flesch-Kincaid grade-level formula, although items are read to the examinees, not read by them. All ILK items were reported to be below the 10th grade level, with the overall reading level of the ILK reported as 5th grade equivalent.

Next, the ILK items were reviewed by a panel of nine, mostly forensic experts and lawyers, for possible bias due to gender, race, religion, or other factors, and problems due to item format. Revisions were made to several items.

Finally, the ILK was administered to 105 psychiatric and forensic patients under either honest or fake-bad instructions. Data was used to examine item-scale correlation and item variability, and two items were identified for possible deletion. However, other considerations led to the items being retained. Some additional changes in item language were also made.

The steps described above help ensure the usability of the ILK and provide evidence of its fairness when used with different populations. However, no formal content validity steps appear to have been taken: The authors do not report factor analyses to demonstrate that the ILK measures a single, coherent construct—or alternatively, a set of related or discreet constructs. Although the ILK addresses a wide variety of content, there does not appear to have been an independent review by legal scholars. Finally, apart from comprehensiveness, it is not clear if different topic areas, which may be more or less crucial to a factual understanding of the proceedings, are weighted optimally. However, these issues apply in greater force to a true psychometric test of legal knowledge, a role the authors do not claim for the ILK.

Another issue not discussed, with relevance for the ILK's use as a response style measure, is variance in jurisdictions across the country. One item (59) is keyed incorrectly for New Hampshire. The authors report that attorneys from various states did review the items for such issues, but the review apparently missed at least one problematic item.

Psychometric Qualities

The authors report that ILK has a Cronbach's alpha of .88 and a two-week test-retest reliability of .76. Based on the former figure and the SD for clinical populations, this gives the ILK a standard error of measurement (SEM) of 2.77 raw points, and a 95% confidence interval of about ± 5.5 points.

Several forms of validity evidence are presented. The authors report that the ILK correlates substantially with Trial 2 of the TOMM ($r = .62$), with the Rey Fifteen Item Test (FIT; $r = .60$), and with Reliable Digit Span (RDS; $r = .61$), and that all correlations were significant at the .01 level. However, it should be noted that FIT and TOMM have both been faulted as showing limited sensitivity to feigning (Boone & Lu, 2007; Gervais, Rohling, Green, & Ford, 2004; Grote & Hook, 2007; Nitch & Glassmire, 2007). Although there is evidence that at least one cognitive validity measure (the TOMM) performs better in criminal populations than it has in civil (disability) settings, the FIT and RDS

showed much lower failure rates than some other well-validated cognitive validity tests (Ardolf, Denney, & Houston, 2007; Denney, 2007).

The ILK manual presents data from both simulation and known group studies. Groups of simulators were created from actual community psychiatric inpatients and insanity acquittees. Although a similar comparison was made using college students, this has limited relevance to the prime validity issue: the ILK's ability to distinguish feigners from genuine psychiatric patients, particularly those found to be incompetent to stand trial. A table on page 32 presents means, standard deviations, and Cohen's *ds* to compare simulators with honest responders. Most honest clinical groups produce ILK scores of 50-53; community (non-psychiatric) controls score a bit higher (55-56), community juveniles considerably lower (mean = 49.0), and defendants evaluated to be incompetent scored a mean of 40.6 (SD = 8.53).

In the simulation study, half of community inpatients and insanity acquittees were asked to respond honestly, the other half to fake bad. The manual reports that the ILK was able to separate honest responders from feigners with Cohen's *ds* for community psychiatric inpatients and insanity acquittees of 1.31 and 1.43, respectively. These are large effect sizes.

Validity data from known groups are limited. A group of 60 malingerers was identified from defendants undergoing competency evaluation. The authors report that at least one other validity instrument apart from the ILK was used to establish malingering. The ILK separates identified malingerers from defendants found incompetent in a contemporaneous evaluation with a Cohen's *d* value of 1.62.

Another way to consider validity is through the *likelihood ratio* (LR), which is the *sensitivity* divided by the *false-positive rate*. Stated differently, the LR quantifies how much more likely those with the condition of interest (e.g., feigning) are likely to be detected relative to those without the condition. Using the community psychiatric inpatient and insanity acquittal samples, cutoff scores were examined for utility in separating honest responders from those patients in those same samples instructed to fake bad. At the recommend cutoff score of ≤ 47 , the ILK manual (Table 3.4, p. 19) reports a sensitivity of .76 and a specificity of .79 (false-positive rate = .21). Thus, the LR = $.79 / .21 = 3.76$. This is a respectable level of discrimination. However, the false-positive rate requires further attention.

Most authors of validity tests consider the diagnosis of malingering or feigning a very serious one requiring very strong evidence and conservative decision rules. For example, the authors of the original SIRS set their cutoff rules to keep false-positive rates below 2.5%, although this meant compromising the ability to identify feigners (Rogers, Bagby, & Dickens, 1992). Most guidelines for interpreting below-chance performance

use a 95% confidence interval, hypothetically tolerating a 5% false-positive rate for those who would truly perform at chance level because of profound impairment. For the ILK, the false-positive rate reported at a cutoff score of 47 is 21% for clinical samples. At the lowest cutoff score for which full diagnostic statistics are reported (≤ 36), the false-positive rate remains at 8%. Taken at face value, this suggests that scoring below the ILK normative cutoffs should only be taken as suggestive of feigning, not as a diagnostic finding. Thus, scoring below these cutoff scores would not be equivalent to, for example, failing the SIRS-2 or TOMM.

Users should also be aware that a base rate of .47 was used in calculating the positive and negative predictive power (PPP, NPP) statistics. Although this figure is consistent with recent data (Ardolf, Denney & Houston, 2007) and my own experience, it is much higher than the 15-17% figure cited by Rogers and Bender (2003). Users should be aware that if the base rate in their facility is much different than 50%, the PPP and NPP will be markedly different.

The foregoing material supports the ILK's reliability and validity. However, a closer look at the ILK's validity data raises some serious concerns. The clinical and forensic groups on which they are based appear to be samples of convenience, such as insanity acquittees and actual criminal defendants. Some were found incompetent, others competent. So far so good. The problem is, there is nothing in the manual to indicate the subjects' clinical status at the time of the ILK administration. Among the information not reported are diagnoses, mental-status variables, current symptoms, IQ scores, time in treatment, medication status, and privilege status within the institution. An insanity acquittee is likely to have been restored to competence before being found insane, so there is little reason to believe most would be psychotic or seriously impaired at the time of the ILK administration. Similarly, someone found incompetent would likely be under treatment to restore competence. If the ILK was administered two or four months after the competency determination, there little reason to believe the person was still incompetent, absent documentation to the contrary. Without well-defined comparison groups, there is little evidence to show the ILK can separate feigners from those clinically impaired to the point of incompetence *at the time of the competency evaluation*.

In fact, the data in the manual show the ILK has only modest ability to do so. Seventeen examinees were found to be incompetent and produced a mean ILK score of 40.6 (SD = 8.53). Simulators from two clinical groups produced mean scores of 36.2 and 35.5. This is the most relevant evidence of the ILK's validity reported in the manual, and the contrasts produce Cohen's *ds* of .40 and .48 respectively—much smaller than the figures reported for patients who have not been shown to be seriously impaired or incompetent at the time of the assessment. The difference in mean scores is just at the edge of the ILK's 95% confidence interval based on SEM. Finally, the mean of actual incompetent defendants is well below the ILK's recommended cutoff score of 47. How

can scores of 30-46 be construed as evidence of feigning when they are the range in which most incompetent defendants will score? The false-positive rates reported in the manual are based on clinical groups, not defendants presenting as currently incompetent. This is the appropriate group on which to calculate the false-positive rate, and the data presented in the manual in Table C1 on the small incompetent sample ($n = 17$) shows 82% scored below at or below 47. This is clearly unacceptable and indicates that any useable cutoff score would have to be much lower. Table C1 indicates that one defendant classified as truly incompetent scored 29, another 30—both within the level of chance. In order to achieve a false-positive rate of 5% or less based on the limited normative data, a cutoff score of 28 would be required. However, with such a small sample, estimates of the false-positive rate will have a large band of error. Until a sample of perhaps 150 incompetent defendants, across several test sites, is obtained, users should probably not consider ILK scores as an indication of feigning except when they are significantly below chance.

The known group data also has serious problems. The manual states that both the ILK and other validity instruments were used to classify malingerers *in the majority of cases* (italics added). Thus, it appears the ILK may have been the only instrument in some cases. Although this would seem to be a clear case of criterion contamination, the authors assert otherwise. Further, because the authors wished to be conservative, only ILK scores that were below chance were used in identifying examinees as malingering. The import of this contamination depends on the degree to which other instruments and data contributed to the malingering classification. However, the net effect, unless the ILK contributed to very few classifications, is to differentially select cases with very low (below chance) ILK scores. This will artificially depress the average ILK score of those classified as malingering, and bias effect size estimates.

There is evidence that such biased selection did occur. The known group malingerers scored fully 10-11 points lower than simulators from clinical groups, a difference that is both statistically significant (both $ps < .01$) and large (Cohen's $ds = .92$ and $.98$). Data and effect sizes for these groups are shown in Table 1. Thus, the low mean score of the malingering group, and the large effect sizes generated by contrasting the malingering and clinical groups, cannot be given much weight in assessing the ILK's validity.

Table 1

Descriptive Data and Effect Sizes for the Most Relevant ILK Samples

Group	<i>n</i>	Mean	SD	Cohen's <i>d</i> ^a	Cohen's <i>d</i> ^b
Currently Incompetent	17	40.59	8.53		
Simulators (psych inpatients)	50	36.24	12.75	.40	.96
Simulators (NGRI acquittees)	43	35.49	12.20	.48	.92
Malingering	60	25.03	10.58	1.62	

^aCohen's *d* contrasting with the Currently Incompetent group

^bCohen's *d* contrasting with the Malingering group

The validity problems are magnified for defendants with a history of intellectual impairment. No control group data are presented for individuals with IQs substantially below normal. Since people who feign incompetence may well present as intellectually impaired, and actual intellectual impairment can be a legitimate reason for incompetence, such data could be very useful. However, it might also be that the ILK cannot distinguish between low ability and feigning until below-chance performance is reached. Until such studies are completed, the ILK should be used with caution with persons with significant intellectual limitations, and only significantly below-chance scores interpreted.

User Friendliness and Usefulness

The ILK is easy to use. After a quick orientation in the manual, administration is straightforward. Even after six months, about the only thing that might be forgotten is the need to provide feedback after each answer. Since most forensic examiners will not use each instrument in their toolbox regularly, it is important that they be intuitive so that they can be used infrequently without a high risk of an invalid administration. The ILK meets this criterion. It takes about 15 minutes to administer and about one minute to score. The cost per administration is a little over \$2.00 apart from the initial investment in the item forms and manual. A specimen kit, which includes 25 answer forms, is a very reasonable \$129.

However, the ILK is less versatile and useful than it might be. The first limitation stems from lack of norms with a cognitively impaired sample. The second comes from the authors' decision to develop the ILK primarily as a response-style instrument, and to recommend its use for only this purpose. The ILK could function as one measure of a defendant's factual understanding of the proceedings. It is not without limitations in this role, as a broad true-false measure may not uncover specific deficits that might be cru-

cial. After all, someone has a 50% chance of being correct on the true-false item just by guessing. However, defendants may present in ways that present opportunities to use the ILK as something other than a response-style measure. I recently assessed a defendant who was quite uncooperative in answering most questions, including about the court system, and I believed he was faking. I thought this would be a good chance to try out the ILK, and fully expected him to fail. But he didn't. In fact, he scored 53, which is only a couple points below the mean for adults without a psychiatric history. Although the ILK did not provide evidence of malingering, it did provide evidence that the defendant had adequate knowledge of the court system. The authors hint at interpreting high scores at several points in the manual, but do not endorse it. There is validity evidence pertinent to this use in the manual, in that one incompetent sample scored well below the mean of other groups, but another group did not. These results are complicated by the possibility that those found incompetent were not judged so because of lack of knowledge, but because of irrationality. I would like to see further development of the ILK as an adjunct measure of legal knowledge, as it is the only true-false format competency instrument around, and it does cover a wide range of relevant content.

Summary

The ILK is easy to use, reasonably priced, has good reliability, and is a thoughtfully developed response-style measure in many ways. However, it seems that the authors adopted a liberal conception of validity: that the ILK be able to distinguish those who feign competence from general psychiatric and forensic patients. This is not the primary decision that forensic examiners face: They must determine if a defendant who presents as seriously impaired, cognitively or psychiatrically, is truly incompetent or feigning. Psychiatric patients who present with depression, anxiety, PTSD, or substance abuse—even if requiring inpatient treatment—are not fully matched controls unless they also present as sufficiently impaired to be incompetent to stand trial. At present, the only ILK sample with evidence of serious impairment at the time of the assessment did not score much different from simulators. With an effect size of .40-.48, one might expect an overall classification rate on the order of 60-65%, as opposed to approximately 77-78% suggested in the ILK manual. Perhaps more important, scores from 30 to 46, although below the ILK's recommended cutoff score, are fully consistent with both feigning and actual incompetence. This writer strongly cautions against use of the recommended cutoff score of 47. Because the current sample of incompetent defendants is very small ($n = 17$), any recommendations about norm-based cutoff scores must be very tentative. Although the limited normative data suggest scores of 33 and below are increasingly suggestive of feigning (see Table 2), the author recommends only significantly below-chance scores (< 24) be interpreted as strong evidence until a more substantial sample can be collected. However, based on the simulator data (p. 44), only about 16-20% of malingerers will score this low.

Table 2

Sensitivities, False-positive Rates, and Likelihood Ratios of Different ILK Cutoff Scores

Cutoff Score	Sensitivity	False-Positive Rate	Likelihood Ratio
28	24	<1	>24
29	25	6	4.2
30	28	12	2.3
31	32	12	2.7
32	34	18	1.9
33	38	24	1.6

Note. Sensitivity is based on detection of simulators in the Community Patient and NGRI Acquittee Fake Bad Conditions ($n = 93$). False-Positive Rate is based on Competency Examinees, Incompetent ($n = 17$).

Developers of cognitive validity tests, like the TOMM and WMT, began by identifying tasks that were minimally affected by brain damage or legitimate cognitive disorder. Although I would have thought that basic knowledge of the court system would not be much affected by mental illness, the ILK data suggest either that it is, or that cognitive limitations (e.g., MR) were affecting scores among those deemed incompetent. Since no data were reported about the intellectual functioning or clinical status of subjects, it is not clear which factors are in play. It does seem likely that serious intellectual limitation might result in very low scores on the ILK, but of course, not less-than-chance performance. It would be very desirable to know if psychiatric illness can substantially reduce ILK scores, absent intellectual limitations or gross confusion. It could be that the subjects found incompetent in the ILK data set were, on the average, intellectually limited. If so, the ILK may function much better when intellectual factors can be ruled out.

Standards have been evolving outside of psychology for reporting of the methodology and evaluation of diagnostic tests. The *STARD initiative* <http://www.stard-statement.org/> (Bossuyt, et al., 2003) is an effort to create greater uniformity across professions and promote full disclosure of factors that can affect diagnostic statistics. The ILK manual falls considerably short of these recommendations, although it is not alone in this regard (Rubenzer, 2010).

The ILK potentially offers strong evidence of feigning if below-chance levels of responding are obtained. Users are cautioned that normative interpretation of the ILK, using a cutoff score of ≤ 47 , can be expected to yield high rates of false-positive errors—as high as 82% based on the small sample of incompetent examinees. This cannot be justified. However, using significantly below-chance responding as a criterion means that the great majority (82%) of feigners will go undetected on the ILK. Users must be prepared to explain that passing the malingering test they used actually does not mean that the person is responding honestly, and why they chose to use a test with such low sensitivity for the condition of interest. Given these observations, it is hard to recommend the ILK until more favorable, quality data are presented. Examiners that use the ILK need to fully understand the issues discussed in this review or risk seriously misjudging the people they evaluate and misleading the trier-of-fact.

References

- Ardolf, B. R., Denney, R. L., & Houston, C. M. (2007). Base rates of negative response bias and malingered neurocognitive dysfunction among criminal defendants referred for neuropsychological evaluation. *Clinical Neuropsychologist, 21*(6), 899-916.
- Ben-Porath, Y., & Tellegan, A. (2008). MMPI-2 RF™ (Minnesota Multiphasic Personality Inventory-2 Restructured Form™). Pearson Assessment Systems. Minneapolis: University of Minnesota Press.
- Berry, D. T., & Schipper, L. J. (2008). Assessment of feigned cognitive impairment using standard neuropsychological tests. In R. Rogers (Ed.), *Clinical assessment of malingering and deception, 3rd Ed.* New York: Guilford.
- Boone, K. B., & Lu, P. H. (2007). Non-forced choice effort measures. In G. Larrabee (Ed.), *Assessment of malingered neuropsychological deficits.* New York: Oxford.
- Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., Lijmer, J. G., Moher, D., Rennie, D., & de Vet, H. C. W. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. *Clinical Chemistry, 49*, 1-6.
- Denney, R. I. (2007). Assessment of malingering in criminal forensic neuropsychological settings. In K. B. Boone (Ed.), *Assessment of feigned cognitive impairment: A neuropsychological perspective.* New York: Guilford Press.
- Gervais, R. O., Rohling, M. L., Green, P., & Ford, W. (2004). A comparison of WMT, CARB, and TOMM failure rates in non-head injury disability claimants. *Archives of Clinical Neuropsychology, 19*(4), 475-487.
- Green, P. (2005). *Green's Word Memory Test user's manual.* Edmonton: Green's Publishing, Inc.
- Grote, C. L., & Hook, J. N. (2007). Forced-choice recognition tests of malingering. In Boone, K. B. (Ed.) *Assessment of feigned cognitive impairment: A neuropsychological perspective.* New York: Guilford Press.
- Miller, H. A. (2001). *M-Fast: Miller Forensic Assessment of Symptoms Test professional manual.* Odessa, FL: Psychological Assessment Resources.
- Musick, J. E., & Otto, R. K. (2010). *Inventory of Legal Knowledge professional manual.* Lutz, FL: PAR, Inc.

- Nitch, S. R., & Glassmire, D. M. (2007). Non-forced choice measures to detect noncredible cognitive performance. In Boone, K. B. (Ed.) *Assessment of feigned cognitive impairment: A neuropsychological perspective*. New York: Guilford Press.
- Otto, R. K., Musick, J. E., & Sherrod, C. B. (2010). *Inventory of Legal Knowledge professional manual*. Lutz, FL: PAR, Inc.
- Rogers, R., Bagby, R. M., & Dickens, S. E. (1992). *Structured Interview of Reported Symptoms (SIRS) and professional manual*. Lutz, FL: Psychological Assessment Resources, Inc.
- Rogers, R., & Bender, S. D. (2003). Evaluation of malingering and deception. In A. M. Goldstein (Ed.), *Handbook of forensic psychology: Volume 11. Forensic psychology* (pp. 109-129). Hoboken, NJ: Wiley.
- Rogers, R., Sewell, K. W., & Gillard, N. D. (2010). *Structured Interview of Reported Symptoms (SIRS) and professional manual*. Lutz, FL: Psychological Assessment Resources, Inc.
- Rubenzler, S. J. (2010). Review of the Structured Inventory of Reported Symptoms-2. *Open Access Journal of Forensic Psychology, 2*, 273-286.
- Slick, D. J, Sherman, E. M. S., & Iverson, G. L. (1999). Diagnostic criteria for malingered neurocognitive dysfunction: Proposed standards for clinical practice and research. *The Clinical Neuropsychologist, 13*(40), 545-561.
- Sweet, J. J, Condit, D. C., & Nelson, N. W. (2008). Feigned amnesia and memory loss. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (3rd ed.). New York: Guilford Press
- Tombaugh, T. N. (1996). *The Test of Memory Malingering (TOMM)*. Toronto, Canada: Multi-Health Systems.