Assessment of Malingering and Falsification: Pushing the Boundaries of Knowledge in Research and Clinical Practice

2

David Faust, David C. Ahern, Ana J. Bridges, and Leslie J. Yonce

Authors' Note:
This is the second of two interrelated chapters that appear in sequence (Chapters 1 and 2). In essence, Chapter 2 is a continuation of Chapter 1 and the two chapters together make up one integrated work. We strongly suggest that the chapters be read in order because the comprehensibility of this chapter depends on familiarity with the contents of Chapter 1.

In Chapter 1 we presented a framework for conceptualizing malingering and identifying key clinical and research issues, in particular the need to focus on ambiguous cases and to work toward reducing ongoing sources of error. Various issues and sources of error were covered in detail. In this chapter, we extend our discussion of clinical and research issues. In particular, we provide more in depth coverage of pressing research needs, laying out what we believe to be key conceptual components and challenges and making many suggestions we hope might prove fruitful. We end by touching on a series of caveats for clinical and forensic practice and research.

The third section of Table 1.1 in Chapter 1 (see page 21) lists additional factors that contribute to false-negative and false-positive errors and sets forth what we believe to be high priorities for continued or concentrated research efforts. We will cover these entries in order, some of which are broadly recognized but others of which have been less completely or minimally described.

Mixed Presentations: Injured and Malingering

A litigant should not have to qualify for sainthood to be compensated fairly for genuine injury. Litigants do not find themselves in a legal system that approaches moral and functional perfection, in which virtually all experts perform nearly flawless evaluations and proceed with unwavering objectivity and fairness, thereby minimizing concerns that symptom complaints or test performances reflecting true injury will not be given their just due. Despite what can be the impressive moral character of litigants, excellent mechanisms for resolving legal disputes, and proficient experts who strive for fairness, we are fallible

D. Faust (✉)
Department of Psychology, University of Rhode Island, Kingston, RI 02881, USA

Department of Psychiatry and Human Behavior, Alpert Medical School of Brown University, Kingston, Providence, RI 02912, USA
e-mail: faust@uri.edu

D.C. Ahern
Providence Veterans Affairs Medical Center, Alpert Medical School of Brown University, Providence, AR 02912, USA

A.J. Bridges
Department of Psychology, University of Arkansas, Fayetteville, AR 72701, USA

L.J. Yonce
Minneapolis, MN, 55414, USA

C.R. Reynolds and A.M. Horton, Jr. (eds.), *Detection of Malingering during Head Injury Litigation*,
DOI 10.1007/978-1-4614-0442-2_2,

individuals in fallible systems that sometimes misstep and fall far from the ideal. Given these realities, we require tolerance of and adjustment for the human condition to maximize just outcomes. This is not to suggest a person should be compensated for a feigned injury. Furthermore, in cases of genuine injury, a good argument can be made for detracting something from the level of compensation when elements of exaggeration are present. Such elements can have a deleterious impact on the legal system and society and, among other things, we wish to deter these sorts of behaviors. However, there are compelling reasons to argue for reasonable proportionality between the presence, extent, and type of falsification and the adverse consequences that ought to result. For example, a hardened criminal who falsely accuses a therapist of depraved behavior simply to receive a financial payoff can hardly be grouped with someone who has suffered a severe injury and embellishes just a little when evaluated by a defense expert who has a well-deserved reputation for underestimating loss. It is for these and other reasons that identifying exaggeration or falsification, in and of itself, may fall well short of providing adequate information for appraising a litigant and why neuropsychologists need to be concerned about joint presentations, especially the co-occurrence of malingering and injury.

If we do not accept the extreme position that *any* degree of falsification should disqualify individuals from *all* compensation, then enhancing knowledge of joint presentations may be recognized as something that could be of great practical value and should be given high research priority. It is precisely because injury status and malingering have partial independence from one another that determining whether someone is malingering often will not resolve questions about the presence or extent of injury. Nevertheless, laypersons may tend to see the categories as mutually exclusive (and sometimes are inappropriately encouraged to do so by experts), which is one reason that research on mixed presentations seems so important. The contribution of forensic experts is proportionate to the degree to which their knowledge exceeds the ken of laypersons on critical issues germane to legal disputes, and hence advances in research knowledge and clinical practice in this area could provide major benefits.

We wish to be quite clear that we are not endorsing or excusing embellishment or falsification, but we also think it is puritanical, categorically unwise, counterproductive, and often simply unfair to lump all such acts together. Rather, many such actions fall within the range of normal human failings and may be elicited, as Rogers (1990a, 1990b) noted more than 2 decades ago, by contextual factors. To illustrate, take an item on the MMPI-2 (Butcher et al. 2001) which asks whether one believes that most people will lie to avoid problems. Among contemporary samples of patient and nonpatient groups, between 50 and 60% of individuals responded affirmatively. For an item asking whether the respondent has pretended to be ill to avoid some responsibility, between 59 and 68% of the groups answered affirmatively. Should a neuropsychologist take a rigid stance about such matters, we suggest that he or she might be administered the "test" we have designed and present in Table 2.1. The first two items are modeled after questions from the MMPI-2. The table is intended to bring home the point that not all forms of impression management are equivalent or a basis for denying all compensation for true injury. It might be argued that the entries in Table 2.1 are absurd or pedestrian, which is exactly as intended.

Especially considering our adversarial system in which the attorney is expected to put on the best case possible, plaintiffs who do not exhibit an iota of impression management, overstatement, or exaggeration (i.e., present nothing but injury with a complete absence of spin) are almost certainly the exception, and in many other cases individuals who are clearly exaggerating or falsifying have also suffered some degree of injury. The great majority of cases likely falls between the extremes and involves some combination of injury and impression management or exaggeration. The frequency of such mixed cases has profound implications for mental health professionals involved in legal cases and for researchers. These cases create assessment challenges

Table 2.1 The Malingering/Credibility Test for Experts

Items
I have never told people I was sick in order to avoid some activity I didn't care to do. (Faking illness for self-gain)
I would never avoid paying every last cent in taxes I legitimately owe even if I were positive I would get away with it. (Stealing; avoidance of social responsibility)
I have never taken something like a bar of soap, a small bottle of shampoo, a towel, or a hanger from any hotel room any time in my life. (Stealing)
I have never exaggerated any of my accomplishments or qualities, e.g., my grades, how well I handled some situation, how considerate I was of others, my work performance, etc. (Exaggerating positive qualities; covering up negative qualities)
If I were stopped going well over the speed limit and a police officer admitted the radar gun was broken, I would still report my speed to the best of my ability. (Trying to get around the law)
If a bank's credit officer was interviewing me for a loan I desperately needed, I would never say something good about that person or the bank if I didn't feel it 100%. (Lying to others to influence their reactions)
When my parents asked me what I was going to do when I said I was going out, I told them the complete truth every time. (Lying to others; manipulating others)
When people ask me about my history, I divulge everything, no matter how bad or embarrassing it might be. (Providing a misleading history; not admitting to personal shortcomings)
On first dates, I never tried to create an impression that was even a little more positive than was truly accurate. (Manipulating others for personal gain)
When I was interviewed for graduate school, I was completely frank in responding to all interview questions and made no effort to emphasize my strengths and downplay my weaknesses. (Misleading others for self-gain)
Interpretive Guides
Given the low sensitivity of the test, a negative answer to any item raises a strong suspicion of falsification (simulation) and doubt about all results. Conversely, any positive response demonstrates an unwillingness to admit to personal shortcomings (dissimulation). If manipulation of results is found, the expert is subject to penalty, such as forfeiture of all expert fees earned over the last 5 years.

and critical scientific needs that have been grossly understudied. The fundamental scientific agenda is to devise ways to separate out, to the extent possible, legitimate injury attributable to the event in question from pseudoinjury or false elements, and thereby deliver useable and effective tools for clinicians. The fundamental task and moral agenda for the trier of fact is to try to sort out these legitimate and nonlegitimate elements, and to then apply what has been discerned to deciding liability and damage issues. For the trier of fact, undertaking both factual and moral determinations is congruent with assigned roles because, after all, the normative justification for the legal system is fair dispute resolution.

Whether or not technically appropriate, the manner in which a judge or juror sorts through the litigant's credibility in the area of damages may have a decisive effect on all major elements of the case, including liability. The spillover to liability issues may occur because such determinations often depend largely on the plaintiff's description given the absence or ambiguity of corroborating evidence. For example, the plaintiff may state that she slipped on a patch of ice and not simply over her own two feet, or that some power tool that was supposed to shut off under certain circumstances failed to do so.[1] In the area of damages, many self-reports or symptom complaints (e.g., trouble sleeping) cannot be independently verified. Thus, in general, a plaintiff whose credibility is viewed as questionable or poor may be compensated well below fair value because subjective complaints that cannot be verified objectively are not believed, or may not be compensated at all, no matter the merits of the case and the occurrence of genuine injury.

These matters should concern experts because it may be their results and testimony that help to sway jurors. Obviously, valid conclusions can

[1] To avoid the cumbersome "he or she" or "his or her," we will alternate back and forth when referring to gender.

foster just resolution of cases and errors can move outcomes in the wrong direction. More so, by identifying and explaining mixed presentations and subtler distinctions or combinations, experts may be able to correct overly polarized views of credibility that conceptualize the matter as all or none or that tend to place credibility and injury in opposition to one another. Of course, if the field fails to develop the needed scientific knowledge on mixed presentations or experts adopt overly narrow views of such matters themselves, what is being offered in this domain may do little to enhance the average layperson's understanding. It is because mixed presentations likely occur with regularity, judgments of credibility have such a powerful impact on cases, misconceptualization in this area might be common, and there is so little direct scientific research on the topic that we consider it a pressing research need. For example, as we will discuss, accurate determinations of base rates and the proficiency of detection methods will likely depend heavily on accounting for mixed presentations, and what we do find may show that current beliefs are often off by a wide margin.

Variations in Conjoint Presentations

We ask readers to look back at Figs. 1.2 and 1.3 on pages 15 and 17 of Chapter 1, respectively. Figure 1.2 subdivides groups along a series of dimensions, with the third line representing litigants who undergo neuropsychological evaluation. Of those individuals, some will have brain injury and some will not, and within each of those two subgroups, some will malinger or exaggerate and some will not. As carried over into Fig. 1.3, the possible combinations of these two conditions or dimensions result in four admittedly simplified categories: not injured and not malingering (I–/M–), not injured and malingering (I–/M+), injured and not malingering (I+/M–), and both injured and malingering (I+/M+). Although of secondary importance for the moment, the cases within each category can be subdivided into those that can be identified definitively or nearly definitively (D/ND) and those that are more difficult to identify or are more ambiguous (AMB). Given the particular difficulty they present, the AMB cases are of greater research interest than cases we are already able to identify accurately (a capability often attributable to advances researchers have achieved).

The lower section of Fig. 1.3 sets forth the four possible combinations of accurate and inaccurate decisions for each category and its associated standing on the dual dimensions of injury and malingering. For example, for individuals who are neither malingering nor injured, judgments about both dimensions may be correct (Acc/Acc), they may both be incorrect (Inacc/Inacc), or judgments about either injury or malingering may be incorrect. For those less familiar with the terminology adopted in Fig. 1.3, VN (valid negative) represents an accurate judgment that a condition is absent, VP (valid positive) an accurate judgment that a condition is present, FN (false negative) an inaccurate judgment that a condition is absent, and FP (false positive) an inaccurate judgment that a condition is present. Unlike simple dichotomous choices in which random selection yields a 50% accuracy rate, given the four possibilities, chance level is only 25% (and the corresponding error rate 75%). Thus, the need for help in decision making is magnified when dual identifications are at issue.

Arguably, some types of errors may be more harmful than others, partly depending on the setting or context of decision making. For example, in a criminal context, an individual who does not have a condition that may compromise judgment or impulse control and who is not malingering, and yet is mistakenly identified as a falsifier, may be unjustly denied release. In a clinical context, this error may not cause nearly as much harm to someone who is not injured as it would to a person who is injured and is falsely identified as a malingerer.

In addition, judgments on one dimension may interact with judgments about the other dimension and do so in an unorthodox manner. In Fig. 1.3, some combinations of the injury/malingering categories and decision accuracy status are bounded by different shapes, each of which identifies different possible interactions. Some

errors on one dimension increase the probability that judgments about the other dimension will also be *incorrect*, but other errors increase the probability that judgments about the other dimension will be *correct*. For some combinations *correct* judgments about the first dimension increase the probability that judgments about the other dimension will be *incorrect*. At others times, the accuracy of a decision on either of the two dimensions is unlikely to affect accuracy on the other dimension.

For example, take the third column, I+/M–, and the entry within it enclosed by a rectangle, FN/FP. This tells us it is quite possible that if one makes a false-negative error when appraising injury, the risk of a false-positive error when appraising malingering is elevated. In this case, abnormal test results are not believed to be associated with true injury or perhaps do not seem to fit with expectations for head injury, and hence the odds of falsely identifying them as a product of malingering are likely to increase. Here, error leads to more error. In contrast, look at the first column, I–/M–, and the entry enclosed by an oval, FP/VN. It is quite possible that if one makes a false-positive error in identifying injury, there is an increased likelihood that the absence of malingering will be identified correctly. Here, the first error may decrease the frequency with which the second type of error is made. Ironically, even correct judgments about one dimension may increase the chances of error on the other dimension. For example, consider the fourth column, I+/M+ and the entry labeled, VP/FN (in a hexagon). Here, the correct identification of injury may lead to more frequent false-negative errors when appraising malingering. Thus, we may have incorrect judgments compounding or counteracting other potential errors, or even correct judgments on one dimension leading to greater error rates on the other dimension. Although other and more complex forms of interrelationships between correct and incorrect judgments may well occur, to our knowledge not even such rudimentary and sometimes paradoxical relationships have been subjected to needed investigation. It is additionally disconcerting to think that these same sorts of interactions between correct and incorrect judgments may be passed along to jurors or reinforce their own misconceptions. Consequently, rather than helping to sort through potential confusion, experts may at times compound misunderstandings. It will take high quality science to begin sorting through these complexities.

Conjoint Presentations: Limits in Knowledge and Potential Consequences

We seem to know remarkably little about joint presentations, including such basics as their relative frequency, the resultant distributions of scores on standard tests and effort tests, and the accuracy with which they are identified. Differences in hypothetical expected test results for our four joint injury/malingering presentations are illustrated in Fig. 2.1. This example assumes that a mild traumatic brain injury (mTBI) is at issue. For purposes of simplification, the figure illustrates expected standing on only a single test of ability, but one that uses a forced-choice format. Certainly things become far more complex when multiple assessment devices are used and a much larger database has been gathered, and should the differential include multiple plausible possibilities alone or in combination, such as head injury plus history of alcohol abuse and mood disorder. In Fig. 2.1, better scores are represented to the left and poorer scores to the right. As illustrated, scores for each group may extend over all or part of the normal range, the abnormal range, or a range that falls below chance performance. Shaded areas represent ranges for which determinations of true status (with some exceptions) often prove most difficult for each I/M presentation or tend to cause the most adverse consequences.

The I–/M– group generally obtains normal scores although, as is common on many tests, a relatively small subset falls in the abnormal range. That small subset risks being misdiagnosed as either injured or malingering.

Assuming the capacity being evaluated is often affected in a subset of individuals with mTBI who do not make good recoveries, the I+/M– group obtains lower scores than the normal

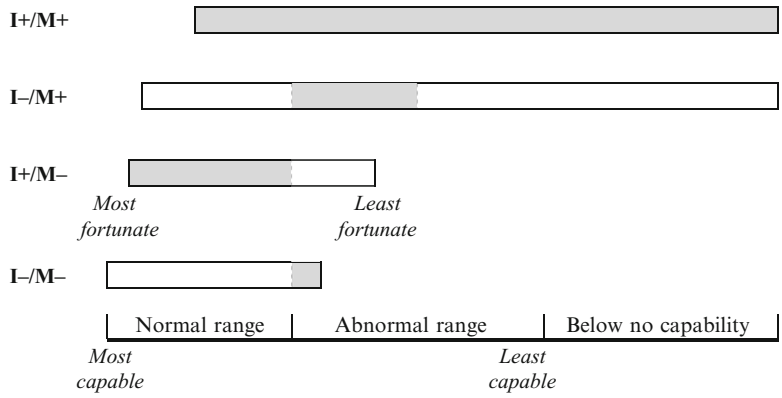

Fig. 2.1 Performance on a quantitative indicator of mild brain injury showing areas of concern (*shaded*) for identifying injured/malingering subgroups

group, a greater frequency of abnormal results, and a broader range of scores. There will be considerable overlap across the groups, and error is most likely to occur when a normal individual performs in the abnormal range or an injured individual performs in the normal range. We understand that deficits following mild head injury can be selective. However, for purposes of this example it should be assumed that some members of the head-injured group have suffered chronic loss in functioning, and for those individuals a normal score represents a false-negative finding when considered relative to overall status. (Alternatively, one could assume this is a cumulative index composed of tests of demonstrated sensitivity to the effects of head injury, all of which use a forced-choice format. However, such details do not change the fundamental situation and are unimportant for the illustrative intent of Fig. 2.1.)

The I–/M+ group may well demonstrate a very wide range of performance, with some individuals, perhaps those who are high functioning and rather selective when feigning deficit, performing somewhat below their baseline but still well within normal limits. The shaded area of the bar for this group falls at a level sufficiently deviant to suggest abnormality but not so extreme as to suggest, or strongly suggest, malingering.

Finally, the I+/M+ group, both having true injury and deliberately not performing up to their capability, almost must obtain scores that on average fall at a lower level than the I+/M– group. Additionally, they are likely to show a very wide distribution of scores. However, unlike the other three groups, with somewhat narrow bands within which determinations may cause the greatest problems, for the I+/M+ group the entire range of performance is of concern. This is because even a relatively strong performance does not rule out a drop from baseline functioning, and even the poorest score does not rule out a degree of true injury. In fact, it is wrong to conclude either that there is no true injury or that depressed scores are due to malingering. Additionally, recognizing that some component of diminished performance may be or clearly is due to inadequate effort is not necessarily that helpful, and making finer discriminations is highly desirable. For example, if in one case we are dealing with 90% malingering and 10% true injury, we probably do not want to group it with another case in which we are dealing with 90% injury and 10% exaggeration, particularly if, in the latter case, even subtle or minor deficits have created major alterations in functioning (e.g., an airline pilot who cannot return to work).

It is natural to ask what scientific knowledge base could assist in differentiating between malingering *or* injured vs. malingering *and* injured and, in the case of joint presentations, in appraising proportionality. As a profession, we might throw up our hands and say it is not possible. However, there are certainly instances in other domains in

which levels of over- or underreporting are measured and the attempt is made, at least if neither is too extreme, to apply corrective measures or adjustments (e.g., the MMPI-2 K Scale). More so, why give up on the challenge before we have really started, especially given the serious implications of incomplete knowledge in this area? Until we know the base rates for mixed presentations and our success in identifying them, estimates that have otherwise been provided for the frequency of malingering and the accuracy of measures are like solving for X with not one but two elements missing – it cannot be done. The consequence is to render assertions about how well we do in this domain and the degree of accuracy that measures achieve, as well as our attempts to incorporate base rates into decision procedures, as crude guesswork or worse.

Let us illustrate the basis for these concerns. Suppose an author claims that the base rate for malingering is 25% in a forensic population and that one or another malingering measure is accurate in 70% of cases. Aside from previously raised concerns about the soundness of such estimates and the overriding need to determine base rates for appropriate subgroups, such figures are likely to disregard mixed presentations. First, we do not know what proportion of this 25% also has nontrivial injuries. Second, given the propensity to treat malingering as merely a present/absent phenomenon and to set high thresholds for identifying its presence, we also do not know what percentage of individuals considered nonmalingerers (the remaining 75%) also show mixed features. Third, assuming interest in identifying mixed presentations, we have virtually no evidence on this matter relating either to base rates or the accuracy of identification. If both malingering and injury status are highly relevant concerns, and as we have argued they often are, it is not clear just how far the available information gets us, especially when one starts to explore the range of possible subgroup frequencies.

To further clarify the possible impact of mixed presentations and their occurrence rates, assume in some setting the frequency of malingering is 50%. Assume further that within this group, those that are malingering only and are *not* also injured (M+/I–) can be identified with 80% accuracy. In contrast, those that are malingering *and* are injured (M+/I+) are rarely identified correctly, with accuracy rates falling only at 10%. Given these assumptions, it is possible to examine what occurs as the base rate for the M+/I+ group shifts. (We sometimes alter the order in which malingering and injury status are listed solely for expository purposes, but this change is not meant to convey any substantive difference.)

Suppose that within the overall group of malingerers, 90% are M+/I– cases (are not also injured) and 10% are M+/I+ cases (are also injured). Projecting across 100 cases, the 90 M+/I– cases are identified with 80% accuracy, resulting in 90 × .80, or 72 correct identifications; and the 10 M+/I+ cases are identified with 10% accuracy, resulting in 10 × .10, or 1 additional correct identification. The combined result is 72 + 1 = 73 correct identifications, or a 73% accuracy rate. In comparison to the 80% accuracy rate for identifying malingering alone, the combined rate of 73% is not quite as good but it is still well above chance level. However, what if 40% of the overall group of malingerers are M+/I+ cases (i.e., they also have significant injuries)? Under such circumstances, 80% of the 60 M+/I– are classified accurately, resulting in 48 correct identifications, but only 10% of the 40 I+/M+ cases are classified accurately, resulting in 4 correct identifications. The combined total is 52 correct identifications, or a 52% accuracy rate, which is just about at chance level. To the extent the base rate for M+/I+ cases increases, the situation only deteriorates further. Suppose the overall group is composed of 80% M+/I+ cases. Given this base rate, 16 of the M+/I– cases are classified correctly (20 cases × .80) and 8 of the M+/I+ cases (80 cases × .10), yielding an abysmal overall accuracy rate of 24%. To determine what figures truly apply here one must know such things as the base rate for the M+/I+ group in the setting of interest, and yet our knowledge about these matters is sorely deficient. Although purely anecdotal, one of the authors has asked various plaintiff and defense lawyers what they believe is the most common presentation in personal injury cases. Every one responded that it was probably the I+/M+ group,

or that among those who are truly injured, most show some degree of embellishment or falsification given the nature of the adversarial system.

More generally, our accuracy rates are a combined product of the base rates for the different joint presentations and how well we identify each subgroup. Suppose, for example, that accuracy in identifying three of the four groups is 80% for each group and that the base rates for the three groups are about equal. In contrast, accuracy in identifying the I+/M+ group is only at chance level. If the base rate for the I+/M+ group is 10%, the combined accuracy rate (for all four subgroups) will be about 75%, well above the level possible by playing the base rates. Specifically, across 100 cases, the other three groups will make up 90 cases and the I+/M+ group 10 cases. If the 90 other cases are identified with 80% accuracy, then .80 × 90 cases will be identified correctly, or 72 cases. If the 10 I+/M+ cases are identified at chance level (which, given four possible choices or groups, is 25%), then .25 × 10, or about 2–3 additional cases will be identified correctly. Combining the 72 correct identifications for the other cases and the 2–3 correct identifications for the I+/M+ cases results in 74–75 correct identifications out of 100, or about a 75% accuracy rate. If one plays the base rates, one selects the most frequent outcome. If the other three groups are distributed about equally and hence each occurs about 30% of the time, the accuracy achieved playing the base rates is only 30%.

It is disconcerting to examine what occurs if the base rate for the I+/M+ group increases. For example, if the base rate for the I+/M+ group is 25%, the overall accuracy rate for all four groups combined decreases to about 66%; and if it is as high as 70%, which may not be outlandish in some situations, overall accuracy declines to about 42%. (These frequencies can be derived by following the same steps set forth above for the 10% base rate.) Obviously, shifting assumptions about relative frequencies and accuracy rates for subgroups changes projections, but the disquieting fact is that we really have little idea what figures might apply. The I+/M+ group may be common and perhaps as or more frequent than any of the other subgroups, and yet it is far and away the least studied. This problem is greatly confounded by research designs that emphasize pure or extreme groups and thereby may inadvertently focus on less frequently occurring and nonrepresentative presentations, consequently distorting and limiting the information we obtain. (Additional serious consequences of what we refer to as the *Extreme Group Problem* will be described below.)

Implications of Conjoint Presentations for Clinical Practice and Research

When pondering possible research and clinical approaches for these joint presentations, it quickly becomes clear that we are entering deep waters. This complexity is evident when one considers how simplified the previous discussion has been, focusing, for example, on single variables for exemplars, emphasizing dichotomous categories vs. matters of degree, and not even touching on critical factors arising from distributions as opposed to simple ranges. We will attempt to describe some key issues and leave more detailed discussion for a later planned work.

With two dichotomous possibilities (again simplifying for the moment), one can sometimes go a long way toward decreasing uncertainty or at least resolving pragmatic concerns by making a single correct choice. For example, if one can determine definitively or nearly definitively that someone was not injured, a critical question has been answered and the issue of malingering may become moot. If head injury is ruled out, then whether an individual is malingering might make little difference, and there may be no point in performing testing at all. Suppose the site at which the individual claims to have fallen and sustained a brain injury is monitored with a video camera, the tape is available, and it is clear the head was not impacted at all and that a head injury could not have occurred. There is no point in conducting neuropsychological testing to determine if a brain injury resulted from the event because one already knows it is not the case, and if testing yielded an abnormal result there would have to be some other cause. If no testing is conducted, the hypothetical possibility of insufficient effort on testing becomes moot.

Assuming the definitive video is unavailable (which is almost always the case), other information will still sometimes allow a near-definitive determination. One seeks information that *maximizes diagnostic validity* and *minimizes susceptibility to manipulation*. For example, although the presence and length of posttraumatic amnesia has considerable diagnostic value, self-report of such is highly susceptible to manipulation. It is no revelation to say that the same individual who might purposely underperform on tests might also provide misleading information about alterations in cognitive functioning at the time of the accident. In contrast, other sources of information (e.g., the observations of trained professionals at the scene who are motivated to reach correct conclusions, or information about the individual's actions at the time), while subject to error, are almost certainly less likely to be purposely altered to create false impressions. On occasion, rich sources of dependable information are available that allow one to rule out a head injury with a high level of certainty. We understand that error or manipulation can enter into these matters as well, such as when an individual stages an accident and pretends to be unconscious. However, fallibility and lack of utility should not be conflated, nor are all fallible methods equal because some are far more fallible than are others.

At times, other sources of information, even if minimally susceptible to manipulation, may not help much. If, for example, the occurrence of a mild head injury is in question, a normal CT scan will not get us very far, despite what some individuals might think, as would also be true of a negative EEG when seizure disorder is questioned. No matter how impressive certain technology might appear, when the task is to all but definitively rule out one or the other dichotomous choice, false-negative error rates beyond a very low level are essentially fatal.

Although *ruling out* the injury in question usually resolves major questions, *ruling in* the injury may have surprisingly limited value. Suppose in the case of a small depressed skull fracture, scanning demonstrates a highly localized but unquestionable area of brain damage. This unfortunate occurrence has now been established, but the situation is unlike one in which injury has been ruled out and concerns about malingering often become secondary. Instead, the co-occurrence of malingering can be highly relevant and could even be the major determinant of self-report, test performance, and other manifestations of seeming dysfunction. The potential presence and impact of fabrication remains ambiguous to the extent that evidence about the presence of structural injury, despite perhaps being highly trustworthy, is not sufficiently predictive of functional effects. In many cases, knowledge of structural alteration does not provide a strong basis for predicting or determining functional consequences, especially if injuries are not extreme or occur in certain brain regions, or if one tries to project over longer time intervals. Further, in many cases, the structural changes that can be detected are only rough approximations of brain injury as a whole. As noted previously, in many courtroom cases level of compensation rests mainly on functional changes. It is ironic that functional impairment is so important in so many courtroom cases, that neuropsychological assessment is often geared toward functional assessment and is a potential means for obtaining critical information, but that our measures tend to be both modest or even weak predictors of everyday functioning (see Faust et al. 2011) and susceptible to manipulation. Important progress has been made in the assessment of function and considerable further gains are achievable, but the scientific obstacles will not be easy.

One way to view the appraisal of potential injury is as a task calling for a probability estimate, or as a type of base rate determination. In principle, the probability of injury ranges from 0 to 100%. In some cases, information is available that will all but rule out the possibility of injury and help us to complete the task at hand. Obviously, this will rarely be possible unless the information used to grade the likelihood of injury is cumulatively valid and largely impervious to manipulation. Once the probabilities of injury exceed a certain level, however, we are back to a situation in which the second of the two basic determinations, in this case the occurrence of malingering, retains critical relevance.

The efforts to reach definitive or near definitive judgments about injury and the secondary benefits that can accrue (e.g., no longer having to be particularly concerned about malingering when injury is absent) do not translate well to the appraisal of malingering. For example, even if we arguably could identify malingering with near certainty, it may help little (given current methods) in appraising the presence and severity of genuine injury. Furthermore, the cases in which we can rule out malingering with certainty or near certainty are likely to involve extreme presentations and occur infrequently. Thus, reaching a clear determination about malingering may not help much with the other side of the equation, which is appraising the occurrence and extent of genuine injury.

Research on the conjoint presentations of injury and malingering has certain elements in common with investigations of comorbid conditions, but the parallels are incomplete and fortunately some of the worst methodological conundrums probably do not apply. Neuropsychological disorders or injuries and malingering probably have sufficient qualitative uniqueness that the problem of separating the two and measuring the relative presence of each is not intractable but at least partly solvable. Conceptually, it helps to distinguish the different ways variables may be related to a disorder and to malingering. Variables might: (a) not be valid or predictive in identifying either malingering or the injury in question; (b) show some degree of association with both dimensions; (c) show an association with one of the two but not the other; or (d) show a positive association with one and a negative association with the other. To illustrate these four classes of relationships: (a) certain demographic features might not relate to either malingering or the disorder in question; (b) decreased scores on measures of mental speed may show a similar strength of association with both; (c) anosmia may show a considerably stronger association with genuine disorder than it does with malingering; and (d) willingness to undergo painful medical treatments may show a positive association with injury and a negative association with malingering. This list of potential relationships has a critical omission that is almost always highly relevant: variables that are also associated with other potential conditions or "rule outs." For example, suppose a variable shows a strong association with malingering, minimal association with head injury, but a strong association with, say, sleep disorder, and the latter is among the litigant's complaints or conditions and plausibly associated with the accident. As such, the variable will be of little or no use in separating out malingering and genuine disorder (in this case, sleep dysfunction). It is because litigants often present with a variety of complaints and possible conditions that promising results obtained in studies that exclude more complex presentations or alternative conditions may create very misleading guideposts for success across applied settings. For the moment, however, we will focus on the first four classes of relationships and come back to this last concern later.

Given the problem under discussion – finding effective methods to evaluate the presence and degree of both malingering and injury – variables with no relation to either true injury or malingering are worthless, as are, at some level, variables that have about an equal association with each. (These latter variables can have value for other purposes, for example, if they help in separating one or both of these conditions from other alternatives.) If, in addition to all of the other things we are trying to accomplish through a forensic neuropsychological examination, we are attempting to determine the extent to which results are attributable to malingering and to true injury, a variable with a similar association to both does not move the inquiry forward. It is critically important to distinguish between variables that have a valid association with the conditions and those that help to differentiate the conditions or appraise their relative standing. No matter the degree of validity, if the variable changes to a similar extent when either malingering or injury is present, it will do us no good for these specific purposes. Thus, we seek variables that are both valid and *differentiating*.

Differentiating value is relative, not in the sense that judgments about art may be relative to the

perceiver or constructed, but relative to the task at hand. A variable that assists in distinguishing between, say, malingering vs. head injury might be ineffective in separating malingering from the effects of carbon monoxide exposure. Therefore, the degree of differentiation possible is often specific to the particular conditions or tasks at issue. In many cases, it is a highly variable quality. The need for both validity and differentiating value, and the potential variations in differentiating value for different dimensions and situations make it all but a non sequitur to describe *the* validity of a malingering detection method, especially when phraseology is meant to convey accuracy. Obviously, accuracy is not a global quality and knowledge of validity alone (e.g., association with malingering) is insufficient to make the needed determinations.

In contrast, if a variable has a greater degree of association with only malingering or with only the injury in question or, even better, if the variable is associated with both but the direction of association is reversed, it has differentiating value. We should look, first and foremost, for this latter or final class of variables but, to the extent we come up short, variables with different degrees of association with the two dimensions can certainly also be of value. It should be apparent that studies failing to examine both validity and differentiating value will not suit our pragmatic needs. Furthermore, studies that merely establish an association between a variable and the presence or absence of one factor (either malingering or injury) will not help us here. Even if a variable shows a high association with malingering and consistently differentiates simulators from controls, it does us little or no good because we also need to know whether or how the variable is associated with true injury.

Most current test methods for appraising effort or malingering, to the extent they are effective, tend to work within an important but restricted domain. These methods usually examine for performance below expectation. Even when emphasis is placed on deviation from an expected pattern of results or on deficits in areas in which they are not expected, the final common path for detection of malingering is lowered performance. For example, if someone shows deficits (i.e., low scores) in areas in which one is not supposed to have deficits, this still comes down to a variation on the same theme – performing below expectation, whether this involves much poorer performances than are expected given the injury in question or the presence of deficits where there should be none. There is nothing wrong in principle with this important detection strategy, but it is likely to be ineffectual for detecting other approaches malingerers might use to create misimpressions, such as false attribution or the provision of an exaggerated baseline. Although detection of underperformance may prove sufficient when a falsifier combines these or related strategies with diminished test effort, other fabricators may be sufficiently cagey to avoid gilding the lily and may limit themselves to misreporting. In such cases, most of our routine methods for assessing malingering, especially those restricted to cognitive measures, are likely to fail. Nevertheless, in order to approach the current topic systematically, we will first address underperformance and methods designed to detect it.

The impact of malingering and true disorder on measures, such as test scores, can be additive, distinct, or interactive. To illustrate an *additive* relationship, assume someone's prior ability in an area of memory falls at a Wechsler-type scaled score equivalent of 100. If the individual is only injured (not malingering) the score might fall to 90, if only malingering (not injured) the resultant score might be 85, and if both injured and malingering the score might fall to 80. To describe the relation as additive does not mean *strictly* additive, only that the combination produces a greater impact than either condition alone. Further, the proportionate contributions of one variable will not necessarily hold for another variable. Although malingering might account for, say, 80% of the change on one variable, it might account for a much smaller percentage on another variable.

Relationships for other variables might not be additive but *distinctive*. By distinctive, we mean that whatever the impact of either malingering or true injury on a variable, the other dimension will not exert an influence. For example, although

malingering might reduce performance on a measure of overlearned material by a modest amount, true injury may have little or no impact on that variable. (In our initial description, we were presenting idealized types, but sometimes additive contributions will be so minor that for practical purposes the relation can usually be characterized as distinct without negative consequences.) Finally, *interactive* relationships obviously cannot occur without the presence of both malingering and true disorder.

For these three kinds of relationships, effects will not necessarily be limited to malingering and one particular condition, such as head injury. As follows, malingering may show additive, distinct, or interactive effects with other disorders or conditions that may also be present, some of which may be entirely independent of the event at issue; the same is true for injury or another condition under consideration. Given the range of potential influences on neuropsychological testing and the relative rarity that these influences will be limited to malingering and a specific condition, beyond hypothetical cases or the unusually clean cases that might be selected for inclusion in research studies, one is usually dealing with a complex causal puzzle.

With these preliminaries in place, we can consider the manner in which detection strategies combine with various classes of relationship between malingering, genuine injury, and predictive variables. When the main detection strategy is directed at performance below expected levels, additive relationships between malingering and true injury increase the likelihood that the level of one or the other will be overidentified, and may change these odds markedly. Furthermore, given the predisposition of some diagnosticians to view malingering and true injury as alternative possibilities as opposed to conjoint phenomena (they think too much "or" and too little "and"), the risk of false-negative errors in the identification of true injury or malingering (but not both simultaneously) may also increase. To the extent that true injury as opposed to malingering contributes to lower performance levels and moves one beyond cutting scores, overestimations of the role that malingering plays are likely to become more frequent, sizeable, and serious. Of particular concern, when injury alone is responsible for diminished capability, the least fortunate – those who fall at the negative end of the I+/M− continuum shown in Fig. 2.1 – are most susceptible to false-positive errors in the identification of malingering and false-negative errors in the identification of injury. Figure 2.1 portrays the situation with a mild brain injury, and one might consider how the risk of a false-negative error in identifying injury and a false-positive error in identifying malingering might both increase at somewhat higher grades of injury. At the same time, it is the most unfortunate individuals who have the most to lose (i.e., they have already lost the most and may be most in need). Worries of this sort make the matter of combined presentations especially pressing, and ignoring the issue by concluding, for example, that one knows malingering is present and therefore cannot determine the extent to which true injury is present is not really a default option. Rather, such a position is most likely to cause harm in cases in which there is an especially compelling moral obligation to avoid it.

When using performance below expectation as a detection strategy, one wants to separate the relative contributions (if any) of insufficient effort and true injury. A basic study design would compare the magnitude of impact on test performances for a group that is malingering but not injured, a group that is injured but not malingering, and a group that is both injured and malingering. A related design would start with an injured group and experimentally manipulate level of effort to examine the effects of such variation on results. One could also use designs that keep level of effort constant and use appropriate patient selection to vary level of injury.

Another approach, which might also start with a group that is malingering only and another that is injured only, would search for variables that achieve both validity and differentiating value. Such study designs have been used frequently with the aim of identifying variables that are likely to be altered by either true injury or malingering

but not both. It is even better if one can find variables associated in opposite directions with injury and malingering. These study designs might be further bolstered by introducing groups in which malingering and injury are intermixed to varying degrees to examine the impact of conjoint presentations. For example, working with an injured group, one could experimentally manipulate level of effort. The goal is to determine if certain characteristics help separate the relative contributions of true injury and level of effort, with the long-term aim being the possible development of corrective methods or adjustments. Corrective methods or scales are commonly created for personality tests, and some of them, at least within certain ranges, demonstrate at least modest levels of efficacy. Little effort has been made to develop corrective methods for cognitive tests, perhaps because feigning is too often treated as a dichotomous variable or because measures are mainly designed to detect only grossly inadequate effort, both of which seem to have in common too much "or" vs. "and" thinking.

Most alternative detection strategies depend on some sort of variation in expected performances (although some of these merge into methods aimed at underperformance). For example, one might look for deviation from expected course over time, atypical symptoms such as the copresentation of complaints that usually do not go together (e.g., a report of anosmia but heightened smell sensation at other times), or deviation from expected highs and lows in test scores. Most such strategies depend on identifying outcome variables that are distinctive (minimally overlap across the injured and those who are malingering) and which thereby may provide both validity and differentiating value. Here again, one can compare individuals who are malingering and not injured to a group that is injured and not malingering, and also implement designs in which level of injury varies and level of effort is experimentally manipulated.

The situation is much more complicated when the impact of injury and malingering interact, and the end result may not be lower performance relative to common levels when individuals are malingering but not injured. For example, someone with true injury may not feel the need to alter performance more than a little to achieve adequate recognition of impairment and may be conservative in these efforts rather than risk being viewed as a complete fraud. In other situations, interactions lead to performance below expected levels for either condition alone. We do not mean to play armchair philosophers and only wish to make two points. First, anticipating interactive effects is often very difficult and best determined through formal study. Second, one thing that is fairly certain is that the majority of interactive effects will alter test patterns.

We mention this matter of pattern alteration with trepidation because we fear it could be mistaken for the argument that such determinations should rest on clinical judgment and that the analysis ought to involve the integration of many variables and the attempt to discern complex interrelationships. We are not arguing for either of these positions and believe they are more counterproductive than constructive (see the discussion of interpretive strategies in Chapter 1). Rather, fairly simple and much more psychometrically sound methods can be used to appraise deviation from expectation. Suppose one identified a composite of variables that were more likely to be impacted by head injury than by malingering and another set of variables that were more likely to be impacted by malingering. Within each composite, the results might be calculated for each component variable by measuring distance from expectation (for that variable) and then summing across the variables. The first cumulative index might assist in judging the likelihood that test scores were impacted by head injury and the second the likelihood that scores were impacted by malingering, with a possible supplemental procedure used to estimate the relative contributions of the two indexes. Variation from expectation does not require complex pattern analysis, and simple linear composites of deviation measurements might be quite effective and easily quantifiable, thereby reducing dependence on subjective judgment.

Obstacles Created by Inter- and Intraindividual Variation and the Need for Baseline Measurement

Even if methods are used that enhance psychometric quality (such as linear composites to increase reliability), approaches emphasizing performance below expected level or deviation from expected performance patterns will almost surely fall far short of their true promise if limited to contemporaneous measurement. Absent sufficient information about baseline functioning, *inter*- and *intra*individual variation tend to overwhelm disorder-specific effects, which is a major problem not only in malingering detection but for almost any form of pattern analysis in neuropsychology, especially approaches emphasizing more than rudimentary configurations.

As is well known, *inter*individual variation refers to distribution in some parameter or characteristic across groups. Even within normal samples or those with no known pathology, the range in proficiency or test scores across individuals is often extreme. For example, the difference between, say, a 40-year-old who obtains a borderline vs. a very superior score on the WAIS-IV Information subtest may be 3 correct answers vs. 25 correct answers (Wechsler, 2008). *Intra*individual variation refers to an individual's range or "scatter" in scores across areas and is frequently much greater than is commonly assumed. As has been known for decades, and as recent research has reemphasized (e.g., Binder, Iverson, & Brooks, 2009; Brooks, Iverson, Sherman, & Holdnack, 2009; Brooks, Strauss, Sherman, Iverson, & Slick, 2009; Dumont & Willis, 1995; Schretlen, Munro, Anthony, & Pearlson, 2003), an individual will often show large differences in performance across areas, especially as the number of administered tests expands. For example, even in circumstances that, if anything, tended to reduce intraindividual variation (e.g., co-norming, a modest vs. high number of tests), Schretlen et al. (2003) obtained a mean difference between an individual's highest and lowest scores of 3.4 standard deviations (SD), with 20% of the sample obtaining differences of 4 SD or more.

By *disorder-specific effects* we mean true differential impact of diseases or conditions on functions. Suppose, for example, that moderate head injury causes a mean decline of 1.0 SD in recent visual memory, 0.5 SD in delayed verbal recall, and 0.1 SD or less in some area of overlearned factual knowledge. Disorder-specific effects have a true magnitude, but measurements of these effects are approximations with varying error terms that, unfortunately, are often disturbingly large for reasons we will touch on momentarily. Different conditions or disorders are likely to have partly or largely overlapping disorder-specific effects, making the term somewhat of an oxymoron. However, what we are referring to is the adverse impact of the condition or event in question and not unique effects in relation to all other conditions and disorders that can alter neuropsychological functioning.

If one only needs to distinguish between normality and a single disorder, one does not worry about overlap with other disorders; but when one must discriminate among possible disorders or combinations of disorders – probably the more usual situation in neuropsychology – overlap in the effects produced by disorders can become a major concern. As the number of possible conditions or disorders that must be considered (e.g., are viable possibilities) expands and as the degree of overlap increases, distinctions become more difficult. This is one reason it may be considerably more difficult to sort out malingering and injury than malingering vs. normality, especially when litigants may also be impacted by multiple other conditions or factors. The difference between a disorder-specific effect and normality, on average, is greater or far greater than the difference between two disorders and conditions that can both impact neuropsychological functioning. For example, although a moderate to severe head injury may cause an average difference of 1.5 SD between those with and without such injury in a certain area of functioning, the contrast may be much smaller between those so injured and those feigning deficit and may run in the reverse direction (e.g., performance averaging 0.5 SD lower for the malingering group).

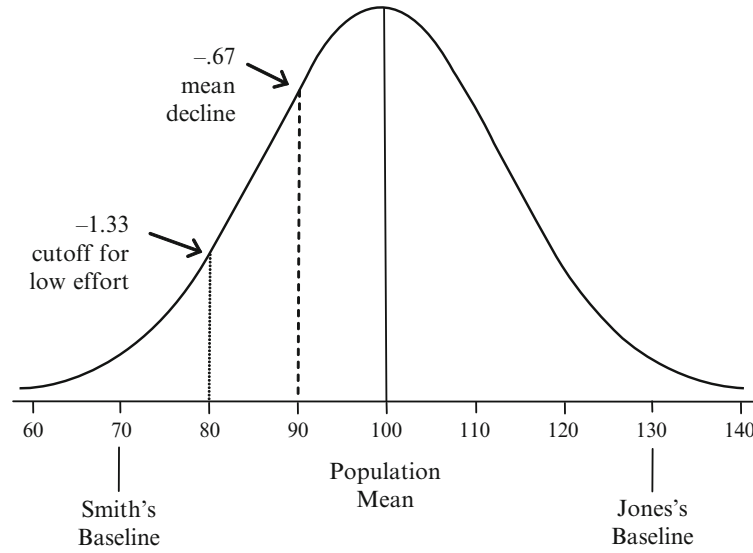

Fig. 2.2 *Inter*individual variation vs. disorder-specific decrements, using mental speed as an example

Given their relative magnitudes, the impact of inter- and intraindividual variation on test performance often overwhelms the influence of disorder-specific factors, especially when one is not so concerned with separating normal individuals from others but rather must undertake more complex differentials. For example, if testing is normal, there is often little left to do. However, when testing is abnormal, one will commonly be faced with a differential involving both malingering and injury, and often will also need to evaluate the potential impact of additional possibilities (e.g., mood disorder, a history of prior head injury, substance use/abuse). The more subtle or less robust distinguishing features are, then to the extent error enters into the analysis due to inter- and intraindividual variation, the greater the diagnostic problem and likelihood of getting it wrong. Inter- and intraindividual variation can be viewed almost as error components (like variation within groups when one is comparing across groups in studies) and may exert much greater impact than disorder-specific effects given the methods of analysis that are routine in clinical and forensic practice.

These abstractions can be concretized through graphic representation, starting with somewhat extreme exemplars for the sake of clarity. Figure 2.2 depicts interindividual variation for a specific function, say, mental speed. Presume a test used to measure this function has a mean of 100 and a standard deviation of 15. The pre-event standings of two hypothetical individuals, Smith and Jones, fall toward opposite ends of the performance continuum, but not at pre-injury levels that extend more than 2 standard deviations beyond the mean. Assume that Smith and Jones both experience mild head injuries and are subsequently tested with a battery that includes this measure of mental speed. For the sake of the example, assume the respective neuropsychologists who examine either Smith or Jones do not have knowledge of prior abilities in this area, or that Smith and Jones had never been tested before the accident. Thus, their baselines in Fig. 2.2 represent their true prior capacities, but these are unknown quantities.

Suppose an investigator has studied the effect of comparable head injuries on this measure. When compared to matched controls, the injured group demonstrates a mean decline of 0.67 standard deviations, or about 10 points, and few injured individuals exhibit a decline of more than 20 points. Thus, a score below 80 is sufficiently unusual among the study sample that it is suggested for use as a cutoff for identifying insufficient effort. The exact figures we set forth in this hypothetical are not important as our main intent is illustrative, but they are not unrealistic.

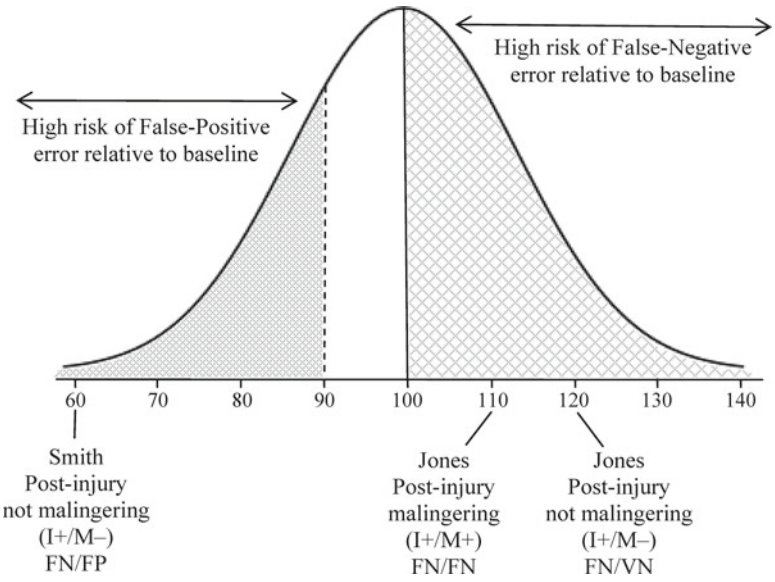

Fig. 2.3 Impact of *inter*individual variability on false-positive (FP) and false-negative (FN) error rates

In studies of standard neuropsychological tests, such results would not be outlandish, nor the suggestion that performances this far below expectation could be considered a possible indicator of falsification. Even should one set the cut-off at a different level, the relative impact on overall error will be similar and will merely change the relative frequencies of false-positive and false-negative error; for example, if one sets a more stringent cutting point to reduce the risk of false-positive error, the rate of false-negative error will increase.

Figure 2.3 demonstrates the damaging effects of interindividual variation on classification accuracy when the obtained cutting score is applied. Smith's injury has produced a drop in his mental speed that translates into a 10-point loss on the measure. He is also making his best effort on testing. Thus, he is injured and not malingering. Nevertheless, he will be identified as putting forth insufficient effort on the measure. He becomes saddled with a false-positive identification of malingering, and his injury might also be missed, an unfortunate false-negative error. We have plotted two possible results for Jones. If Jones is injured and malingering, both are likely to be missed. If Jones is injured and not malingering, then although he at least will probably not be misidentified as a fabricator, his true injury is likely to be missed.

Although Smith and Jones fall at far ends of the continuum, there is a large range of above average (pre-event) ability levels that creates a considerable risk of false-negative error for injury and a large range of below average (pre-event) ability levels that creates a considerable risk of false-positive error for malingering. Furthermore, for those with above average abilities, there is considerable risk that true injury will be missed (with a corresponding increase in the false-positive error rate for those with below average abilities). Therefore, it is not only at the extremes of the continuum for which the risk of error is great when evaluating both malingering and injury status but also for large proportions of the distribution, all due to interindividual variation.

Sociodemographically adjusted norms may help somewhat in reducing the impact of interindividual variation, but often less than is commonly assumed. First, methods that emphasize performance below expected levels on standard tests may not use demographically adjusted scores, hence leading to exactly the sort of problem illustrated in Fig. 2.3. An equal or greater problem starts with both formal and informal methods for estimating prior functioning, but it is too involved to describe fully here. The approach is to adjust expectation for performance based on knowledge of baseline functioning. Thus, if one knows a person was very capable prior to the

accident, expectations for performance (and for identifying performance that raises concerns about cooperation) are modified, as would be the case with someone with low prior capabilities, although in the former case expectations are raised and in the latter case lowered. The limits of using impressionistic methods to estimate prior functioning have been pointed out in the literature, in particular their susceptibility to substantial error (e.g., Faust, Ahern, & Bridges, 2011; Kareken & Williams, 1994; Williams, 1998). With formal methods, whether explicitly stated or not, the main thrust is often the prediction of overall intellectual ability. The problem this creates is that overall ability may not be a strong predictor of other variables assessed during neuropsychological evaluation.

The extent to which neuropsychological assessment enhances or improves upon intellectual assessment alone depends on the independence or nonredundancy of the two methods. If neuropsychological measures correlated too highly with intellectual testing, they would not provide unique or nonredundant information. Hence, neuropsychological measures contribute or add to intellectual testing depending not only on their validity but also their degree of independence from intellectual appraisal and is one of the most basic rationales for the entire enterprise. All else being equal, the greater the independence the greater the extent to which accuracy is increased; that is, the greater the contribution of neuropsychological evaluation to incremental validity. The obverse side of this psychometric fundamental is that attributes that make neuropsychological measures most effective (by maximizing their independence from intellectual testing and hence their contribution to incremental validity) all but ensure that whatever best predicts intellectual testing results will not best predict results on these same neuropsychological measures. If A (a Full Scale IQ score) is minimally related to B (a specific neuropsychological test result), and if C is a strong (or maximal) predictor of A, then C cannot be a strong predictor of B as well. Measures that best predict prior intelligence will often be weak or poor predictors of the neuropsychological tests that are most sensitive to brain injuries, and thus if methods for determining prior functioning mainly address intelligence, they are likely to be unsatisfactory predictors of these neuropsychological measures. Adjustments for sociodemographic factors (in effect, a way of narrowing the comparison group and thereby attempting to better approximate prior functioning) are generally geared toward overall intellectual functioning and therefore are much more effective predictors of that quality than of specific areas of neuropsychological functioning. The end result is that sociodemographic adjustment or other methods that directly or indirectly estimate prior functioning often do not do much to adjust for interindividual (or intraindividual) variation in specific areas of neuropsychological functioning. For example, even if an individual is compared to other individuals with fairly similar levels of intellectual ability, this often does not go that far in assuring similarity in other areas of neuropsychological functioning, especially those areas that are least redundant with intelligence and most sensitive to brain damage or yield the highest levels of incremental validity.

Along related lines, knowing an individual's general level of intellectual ability does not help much with the problems created by intraindividual variability. Figure 2.4 shows Smith having a true pre-event mental speed score of 70 and a true new learning capacity score of 130. Such variation is not unusual (see Schretlen et al., 2003), although in one way we have represented a worst-case scenario because this large intraindividual contrast happens to occur across two areas that are often both affected by head injuries. Suppose again that the cutoff score for poor effort is 80, the injury has adversely affected Smith's mental speed but not his new learning, and that in the latter area Smith is feigning deficit. Due to the combination of baseline functioning and true loss, Smith's mental speed score falls far below the cutoff for malingering. Consequently, the true injury is missed (a false-negative error) and malingering is falsely identified (a false-positive error). Ironically, in the area of new learning, where he has grossly underperformed due to malingering alone, the absence of any effect from

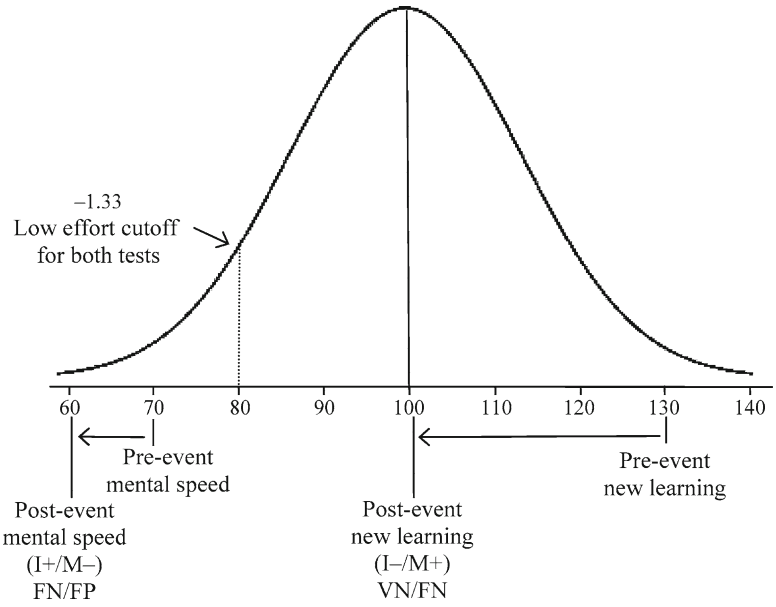

Fig. 2.4 Illustration of impact of *intra*individual vs. disorder-specific decrements

injury is identified correctly (a valid-negative judgment) but malingering is missed (a false-negative error). In fact, Smith would have had to score at least 10 points *above* his pre-injury baseline in mental speed not to be falsely identified as a malingerer, and more than 50 points *below* his true new learning baseline to be detected as malingering in that area. Consequently, although the scenario presented here represents one of the worst problems that can result from intraindividual variation, less extreme occurrences, which are common, can easily lead to errors. More generally, whether using general population norms, impressionistic methods, or formal techniques that are currently available to try to take baseline functioning into account, inter- and intraindividual variation will often overpower or overwhelm disorder-specific effects and lead to frequent errors.

We would go a long way toward eliminating the error component caused by inter- and intraindividual variability with even brief and routine population baseline assessments of neuropsychological functioning. Such information would be invaluable beyond malingering assessment. The tremendous advantages of systematically implemented pre-injury testing are being demonstrated in such areas as sports-related injury, and we hope that a broader lesson can be drawn from such examples. If neuropsychologists can get on board with reducing the length of neuropsychological testing – at least for baseline screening – and increase the use of computer technology, it may eventually be possible to implement broad-based population screening. For example, a 30-min screen administered every decade in widely used health care settings could be of great individual and social benefit and advance the field remarkably.

Neuropsychologists are leery about recommendations to shorten their test batteries, and there are certainly times when truncating assessment due to external pressure rather than best practices is damaging. Even when considering incremental validity of utmost importance, if decisions need to be made about a number of matters and cumulative indices are among the most effective predictors but require fairly lengthy testing, relatively long batteries may be needed to approach a ceiling in effectiveness. However, when the purpose is population-based screening, such lengthy procedures are often impractical, inefficient, and excessive, and if we continue to insist on them the hope of accomplishing this

worthy aim is minimal. It is also shortsighted to believe that reducing length will have a negative long-term economic impact, because almost surely the opposite outcome would result. Imagine if years ago decision makers at IBM agreed that cutting the cost of computers would be bad financial policy. Even if such a policy was financially neutral or negative, which it is not, improvement in patient care should be the determining factor.

By following psychometric principles carefully, a surprising amount could be accomplished via baseline screening in a relatively short period of time. Our techniques for estimating pre-injury functioning are so limited at present that even modest success in baseline assessment would improve our situation considerably. In designing screens with future comparisons in mind, targets might include more frequently occurring conditions which are generally more difficult to identify without a pre/post comparison and for which early identification can reap maximum benefit. For example, the early identification of dementia or its functional consequences, given anticipated improvements in our capacity for helpful intervention, might well be one such target condition. One would likely focus on relatively nonredundant areas in which the greatest changes occur on average, adding items one at a time that make a unique and maximal contribution to incremental validity. When distinguishing between one and another condition is critical, one would focus on variables that have the joint qualities of validity and differentiating value. By measuring and updating appraisal of baseline or pre-event functioning, one can reduce or nearly eliminate two of the biggest sources of error for many forms of neuropsychological evaluation, inter- and intraindividual variation. (We understand, of course, that evaluation of intraindividual variation can serve other critical purposes in neuropsychological evaluation.)

In the meantime, while waiting for these hoped for advances, those wishing to evaluate deviation from expected performance levels have to do their best with what is available. It would seem evident that current approaches for postdicting pre-event functioning using contemporaneous measures and perhaps sociodemographic variables that are aimed at intellectual functioning have limited utility. Studies on these approaches tend to be limited to examining the accuracy with which overall intellectual indices can be determined, and we have noted the limited relation that often holds between general intellectual functioning and the more specific measures that add most to the diagnostic power of neuropsychological assessment. If one wants a sobering look at how poorly such methods seem to work for predicting functioning in specific areas, Schretlen, Buffington, Meyer, and Pearlson's (2005) study provides an instructive example. In essence, one is using a variable (A) to predict intellectual functioning (B) in order to predict functioning in specific areas (C), despite knowing that B often shows limited association with C. The result is a double whammy: predictive power is lost by introducing an additional inferential link (using A to predict B to predict C), and one is using B to predict C despite knowing that the two often do not show a strong association.

If one is going to attempt these approaches, it is much more advantageous to predict C directly from A, by identifying variables that predict functioning in specific areas. For example, with the typical approach, one identifies a combination of variables that correlates with prior Full Scale IQ. These composite variables might achieve a correlation in the 0.50s. One then uses the estimated IQ score to predict, say, the capacity for new visual learning, which may show a correlation of 0.40 with IQ. This obviously degrades predictive capacity severely; if A correlates with B at 0.50, but B correlates with C at 0.40, then the ultimate power to predict C is poor. In contrast, if some combination of variables shows a correlation of, say, 0.30 with C, then by avoiding the extra inferential link this obtained level, although considerably weaker than the original association between the other predictive variables and the IQ score (i.e., between A and B), is still likely to be a stronger predictor.

In principle, so long as one uses neuropsychological measures with satisfactory levels of reliability to assess functioning in particular areas, prior standing in those areas should be as predictable as, say, overall intelligence. For example, if

we have a measure of visual memory with a .85 reliability, that level might not be quite as good as a Full Scale IQ score, but it still gives us a decent chance to identify predictors (or, given the intent to determine prior capacities, postdictors) with at least a modest level of accuracy. In contrast, if measures of specific functions have poor reliability, as is sometimes the case with neuropsychological tests, then the prospects for even modestly effective postdiction are poor. It is almost a given, considering the relatively low correlations between overall intelligence and various neuropsychological tests, as well as the modest to low correlations often obtained among neuropsychological tests themselves, that the best postdictors will often vary from test to test, and sometimes considerably. For example, if new visual learning correlates minimally with finger tapping speed, there is almost no chance that the same set of variables will optimize postdiction of both functions. This is why normative systems that adjust along the same set of sociodemographic dimensions for all tests, even though an important start, have very different degrees of success across measures and will not come close to optimizing postdiction or optimal comparison on a test-by-test basis. Rather, the daunting task, if one were to fully pursue this approach, is to identify the most effective postdictors idiosyncratically, or separately for each measure. Although the uneven success of sociodemographic adjustments often wreaks havoc with pattern analysis, such consequences are rarely mentioned and often seem to operate below the surface or perhaps without neuropsychologists fully realizing what is occurring.

Given all of the limits and complexities involved in using contemporaneous tests to determine prior functioning, one is often better off accessing previously obtained measures. Even here, there is a very limited database on the relation between measures that are commonly used in schools or other settings (e.g., the workplace, the military) and performance on specific neuropsychological measures as opposed to more general measures of intellectual aptitude or academic achievement, and a number of cautions need to be implemented (see Baade & Schoenberg, 2004; Orme, Ree, & Rioux, 2001; Reynolds, 1997; Williams, 1997).

Finally, and perhaps an unsettling thought, the co-occurrence of malingering and injury, which adds considerable complexities to forensic neuropsychological evaluation and research challenges, is only a component of many presentations. Additional factors – some that are causally related to the event in question and some that are not, and some that occupy critical links in the causal chain but that may be subtle, indirect, or multiple steps removed from the original event – all may impede effort or diminish test performance. Consequently, all might impact on predominant methods for assessing malingering, or which rest on performance below expected levels or deviation from expected patterns (e.g., atypical symptoms, atypical course). For example, a car accident may produce orthopedic injuries that cause pain and reduced ability to bear weight. These problems may in turn diminish activity and, when combined with medication side effects, lead to weight gain over the course of months, which produces sleep apnea, which diminishes cognitive functioning and motivation. Many research studies are exercises in oversimplification, something that can be necessary or helpful for certain purposes but may fail to capture the clinician's real world to such an extent that the findings are misleading and result in frequent error when applied directly. In many forensic cases, there may be at least a half dozen causal factors to consider when appraising neuropsychological status and effort, and this is why it is often only the extreme cases (e.g., definitely not injured, or overwhelming injury) that are clear-cut but for which the neuropsychologist's expertise may be least needed.

Mixed Presentations: Some Additional Thoughts

As a starting point, it can be very helpful to sort injury and effort into dichotomous categories. Unless dichotomous classification is performed properly, attempts at greater refinement are

doomed from the start. When either or both are present, a next step can be to determine degree. If both are present and degree can be measured, it may be possible, at least under some circumstances, to adjust measurement of injury in relation to level of effort. For example, it might be possible to develop methods to regress test scores in relation to level of effort. Such corrective methods are likely to be feasible only within a certain range of effort. For example, effort might be so poor that true level of capacity cannot be determined, much as would be the case if an individual responded to every item on a personality questionnaire by providing the deviant answer. It is unrealistic to believe that these difficult determinations can be made routinely and with a high level of accuracy without formal scientific help and properly validated decision rules, and thus the burden falls on researchers to continue the impressive track record of successes and to push the boundaries of knowledge a good deal further. It is as much a mistake to undervalue what has been accomplished as it is to believe that we are all that close to a complete solution.

When examining for performance below expectation, measures specifically designed to assess malingering often produce much greater differences between those who are injured and those who are malingering than do standard neuropsychological tests. The degree of separation these specialized methods create may foster much greater accuracy in dichotomous classification than standard measures achieve. For one, additional factors that diminish effort, such as marked anxiety, may not impact results on the specialized tests very much. Many of the specialized measures are deceptively easy and thus are often insensitive to true injury or conditions that alter results on standard neuropsychological measures. In contrast, many neuropsychological tests are designed to be sensitive to cognitive dysfunction (of numerous potential kinds or causes) and, therefore, scores on them will often be diminished not only by malingering but by dozens of other factors related to true organic or functional malady. Additionally, because specialized tests may create large differences in the first place, even if modest alterations do occur secondary to other variables, overall classifications will frequently remain unaltered. For example, in studies at least, the difference between the injured and malingerers might be about three standard deviations for a specialized test and one standard deviation for a traditional neuropsychological test. Thus, for example, if a moderate-to-severe mood disorder lowers performance on both the malingering measure and a standard neuropsychological measure by about three-fourths of a standard deviation, it will minimally impact the malingering test yet all but obliterate accuracy for the standard test.

It is this positive quality of specialized malingering tests, however, that has yielded a consequence that is likely to greatly compromise their potential for appraising the degree of malingering, especially when it is not extreme, and thereby their potential for assisting in the design and calculation of corrective indices. The separations may be so good that distributions are highly skewed and relatively few errors are required to place an examinee into an extreme class. One end result is that once one exceeds thresholds for poor effort, there may be very little room for variation in test scores. For example, on a popular malingering test, anything below 90% accurate responses can be highly suggestive of poor effort. However, this measure and most others like it are designed to detect extreme departures from good cooperation, and passing them does not mean an individual has necessarily exerted even a modest level of effort, much less a high or optimal level. As a result, scores below cutoff points can be very helpful in identifying poor effort, but scores above them may leave only a few items and a very small range of differing results. Given such truncated ranges for "passing" scores, there is little chance that relative level of effort, or varying degrees of suboptimal effort, can be distinguished or that results could serve a corrective function. Such a limitation, which in no way is intended as a criticism of specialized malingering tests that were designed for other purposes, could be addressed in various ways. For example, one could add branching procedures

that increase the item pool and produce a wider range of possible scores or simply add other measures that may be more effective in assessing relative levels of effort.

The Extreme Group Problem

Both of our chapters emphasize the importance of focusing research efforts on ambiguous cases. The cases with which practitioners need help are not the definitive or near definitive (D/ND) presentations but those involving closer calls, those cases in which it appears as if someone might be malingering but the matter is not that clear-cut. This is the group suspected of malingering, some of whom are indeed malingering and some of whom are not.

It is to the credit of researchers that the percentage of ambiguous cases has steadily declined, although the proportion that remains may be a good deal greater than is sometimes thought due to the more complex determinations that are commonly necessary, such as joint consideration of injury status and malingering. An important additional factor in underestimations of ambiguous cases is what we have labeled the *extreme group problem* (EGP). Stated succinctly, participants in studies overrepresent more clear-cut cases (D/ND malingerers and D/ND nonmalingerers) in comparison to the more subtle or ambiguous presentations that create greater clinical challenges. Overrepresentation of these more extreme cases in turn creates a host of problems, including qualitative and quantitative distortions in research outcomes that frequently undermine generalization and clinical applicability.

Cohen (1988, 1992) subdivided effect sizes into small (0.2), medium (0.5), and large (0.8), not as hard and fast demarcation points but as an interpretive aid. Keeping these proposed classifications in mind, one can ask what is wrong with effect sizes like the following:

4.20	4.57	5.30	8.14
4.23	4.65	5.47	10.24
4.42	4.76	5.74	10.38
4.49	4.90	6.53	13.66

The first and obvious answer is that they are highly implausible and that true differences of this magnitude are almost never obtained in applied psychology. The second and disturbing answer, however, is that they are among the effect sizes reported for scales F and Fb in a meta-analysis of malingering detection with the MMPI-2 (Rogers, Sewell, Martin, & Vitacco, 2003). We are not faulting the authors in any way because they are merely reporting the outcomes of studies, but the fact that effect sizes in this range were obtained in over 10% of the results reported across studies in this meta-analysis is concerning. In another meta-analysis of malingering detection (Vickery, Berry, Inman, Harris, & Orey, 2001), effect sizes for two commonly used measures exceeded 4.00 for 15% of the outcomes. A naïve reader of the literature might take such figures literally and form a grossly overblown impression about diagnostic accuracy in real-life application. Obviously, something is seriously amiss, and what is amiss is the EGP. What these highly implausible effect sizes do not reveal is how much more pervasive the EGP is in areas of psychology and especially in research on malingering detection.

The Rogers et al. (2003) MMPI-2 meta-analysis lists 78 study outcomes or effect sizes for the F Scale and 51 effect sizes for Fb. Figure 2.5 tallies the number and percentage of effect sizes for these two scales that exceeded 1.00. If we take effect sizes of 1.00–1.99 as pushing the boundaries of plausibility, 2.00–2.99 as highly questionable, and 3.00 or greater as very likely implausible, one finds that for the F Scale nearly 7 out of 10 effect sizes (69%) push or exceed the boundaries of plausibility, and more than half (54%) are highly questionable or implausible. For the 51 effect sizes for Fb, 65% (corrected for rounding in Figure 2.5) push or surpass the bounds of plausibility and nearly half are highly questionable or implausible (i.e., ES = 2.00 or greater). The situation is even more extreme for some of the measures included in the Vickery et al. (2001) meta-analysis. Figure 2.5 provides the combined results (designated as *Vickery*) for the Digit Memory Test and the Portland Digit Recognition Test. For these two tests together,

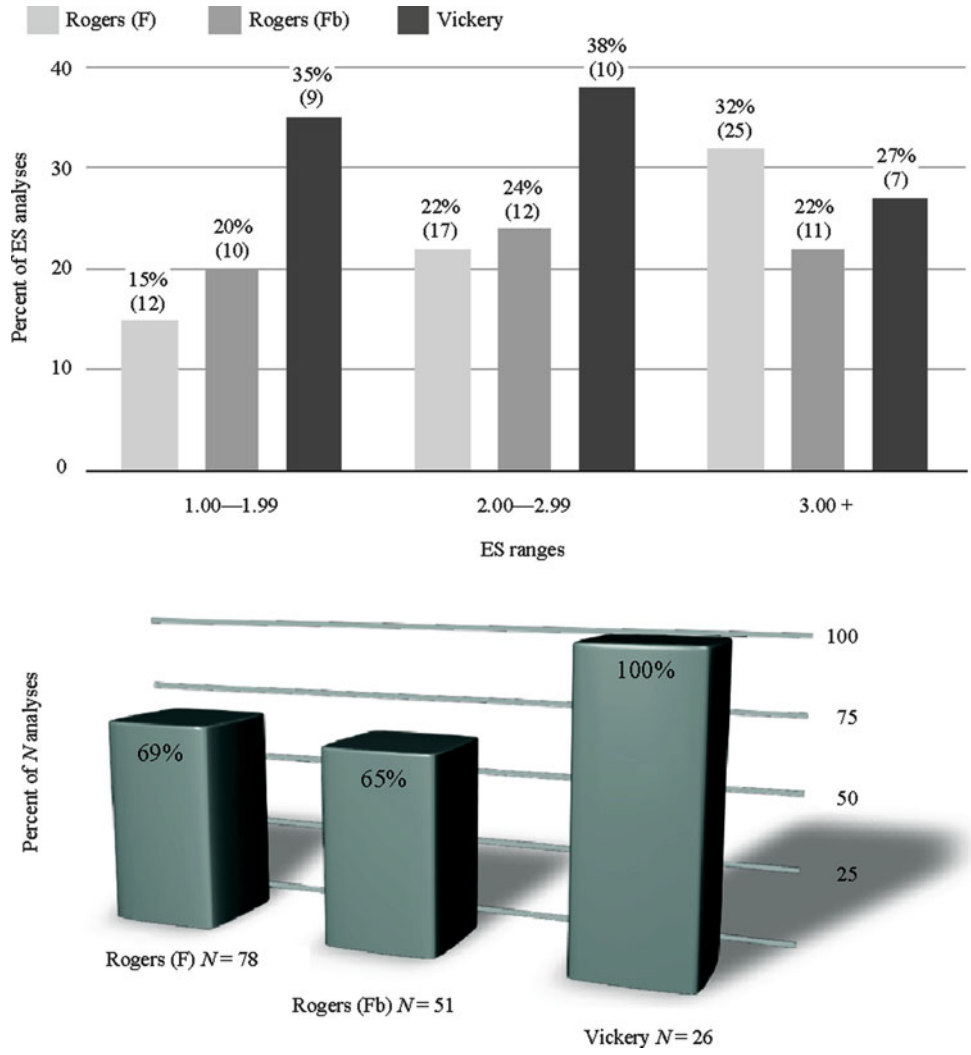

Fig. 2.5 Effect sizes exceeding 1.00 found in meta-analyses by Rogers et al. (2003; MMPI-2 scales F and Fb) and by Vickery et al. (2001) by size category (*top*) and for all ES analyses. (Percentages at top do not total to 100% for the Rogers et al. analyses because ES values below 1.00 are not included here; all of Vickery et al.'s ES values were 1.00 or greater.)

100% of the outcomes push or exceed the boundaries of plausibility, and 65% (i.e., those exceeding 2.00) are highly questionable or implausible.

Our interpretations are not intended to criticize the authors of the meta-analyses or to imply that the measures do not have value. For example, an effect size of, say, 1.0 can reflect a truly robust measure with strong psychometric or predictive properties. Furthermore, even if an effect size is implausibly high, it does not mean that the true or intrinsic validity of the measure is not a contributor to the result. For example, the inflated outcome may be an effect size of 3.75, but the "true" effect size may still be a robust 0.80. (The word *true* is used in quotes because, of course, effect sizes vary with application, and the term is intended to reference the true value for the intended application.) What we do think is abundantly clear is that many of these effect sizes are inflated, often by a sizeable amount, and that the problem is pervasive. This is, we believe, the EGP rearing its ugly head.

Explanation of the EGP

The EGP is a subtle and often underappreciated, but potent, methodological flaw that distorts the outcomes of studies and leads to inflated effect sizes. The degree of inflation can be extreme and lead to gross overestimation of effectiveness in the settings of intended application, such as forensic neuropsychological assessment in civil litigation. The EGP is by no means limited to research on malingering and occurs in numerous other appraisal domains (see Bridges, Faust, & Ahern, 2009; Faust, Bridges, & Ahern, 2009a, 2009b), but its impact seems to be especially pernicious in malingering assessment.

For the moment, we can designate those suspected of malingering as MS, and further subdivide this group into those who are and are not malingering, respectively represented as MS+ (suspected and malingering) and MS− (suspected but not malingering). Presumably, malingering is suspected because of some departure from regularity, such as lower than expected scores on one or more standard neuropsychological measures. In research studies, the EGP is usually produced by the methods used to select both the malingering and the nonmalingering groups. In many cases, selection procedures for both groups result in extreme cases. The malingering group is more extreme than the typical MS+ case, with the difference representing more extreme cases of malingering. The control group's deviation is usually in the other direction, that is, the group is more normal or unremarkable than the MS− group. Thus, both research groups are more extreme than typical cases but in opposing directions: the malingering group is more deviant than the MS+ group and the control group is less deviant (or more normal) than the MS− group.

One might ask why group selection procedures would permit this type of nonrepresentativeness to occur. Understandably, when selecting members for the malingering group, the researcher wants to be fairly certain that group members are malingering. Hence it is common to use fairly stringent inclusionary criteria that require markedly deviant results, such as clearly elevated scores on multiple malingering tests. Conversely, the researcher also seeks reassurance that those in the control group are not malingering, in which case fairly stringent inclusionary criteria might be set in the other direction. Here, one might require clean results on malingering tests and perhaps satisfaction of other criteria as well, such as certain minimal scores on standard tests and status in a group (e.g., nonlitigants) in which incentives to malinger are limited or negligible. Ironically, one is therefore selecting individuals for whom there is little reason to suspect they are *not* malingering (the experimental group) and others for whom there is little reason to suspect they *are* malingering (the control group) to learn how to identify those whom we suspect are malingering and are malingering (the MS+ group) and those we suspect are malingering and are not malingering (the MS− group). If simulation designs are used, similar problems may occur, if for no other reason than because a group of normal individuals who perform their best will often be markedly more normal or intact or neuropsychologically superior in comparison to the MS− group.

Evidence suggests that the EGP, or the magnitude of this methodological flaw, frequently accounts for far more variance in the outcomes of studies than the intrinsic or true quality of tests or assessment methods. If the numerous exceedingly high effect sizes in various meta-analyses are accounted for primarily by the EGP and these effect sizes may be inflated by a factor of two or three (or more), then clearly the EGP is the most influential determinant of outcome. We are obviously in a very bad methodological situation if the worse the design flaw, the better a method performs in studies, especially if the presence or magnitude of the EGP is underappreciated or not recognized. When there is a *positive* association between the degree of methodological flaw and the level of accuracy studies yield, we can be driven further and further from verisimilitude or the correct evaluation of methods.

In addition, if we are comparing different tests or assessment methods and the background studies do not overlap sufficiently, as is very common in malingering detection research, relative merits can be skewed or grossly distorted. Suppose test

A is truly much better than test B. However, suppose further that the studies on test A are minimally saddled with the EGP but a separate set of studies on test B show this problem to a marked degree. As a consequence, accuracy rates or effect sizes generated for test B may seem much more favorable than those generated for test A. As we will later show through example, this sort of situation is not an abstraction, because nonoverlapping studies are common, even in meta-analysis, and have the potential to alter or even reverse rank ordering of efficacy. Consequently, even highly conscientious neuropsychologists who carefully incorporate scientific literature into their practices may be inadvertently led into making poor choices.

To illustrate the potential for distorted ranking due to nonoverlapping studies, we can momentarily turn back to the Rogers et al. (2003) meta-analysis of MMPI-2 malingering scales or indices. As we noted, 78 effect sizes were reported for the F Scale and 51 for the Fb Scale. For almost every study involving Fb, results were also reported for F (50 of 51 analyses), something explained by the predominance of the F Scale as an MMPI-2 malingering indicator. Thus, in 50 cases, the study groups were the same for F and Fb, thereby holding the EGP constant. For example, if a study examined effect sizes for Fb for two groups of subjects, the same two groups were used to calculate effect sizes for F as well. Whatever the magnitude of the EGP in those analyses, it was the same for both groups. There are different ways to summarize the results of the meta-analyses for these same-groups comparisons, and we will focus on a simple indicator for illustrative purposes.

For the same-groups analyses, 78% of the effect sizes exceeded 1.00 for F and 74% for Fb, suggesting that when the EGP is held constant the F scale might be a slightly stronger malingering indicator than Fb. This is not to say that the obtained effect sizes reflect applied performance, but if the EPG is held constant, then the *relative* efficacy of indicators or tests should not be altered or distorted. Of interest, in the 28 nonoverlapping studies – those that examined only F – 54% of the effect sizes exceed 1.00, a considerably lower figure than the 78% obtained in the same-groups analyses. It is likely this lower rate was obtained because, on average, the nonoverlapping studies have less extreme groups in comparison to the same-groups analyses. In the Rogers et al. meta-analysis, summary figures are provided for F and Fb, the former of which is a composite of the 50 same-groups analyses and the 28 nonoverlapping studies. Given the lower effect sizes for the F scale found in the nonoverlapping analyses, the composite figure derived by combining the nonoverlapping analyses and the same-groups analyses is 69%. Thus, in the same-groups studies in which the EGP is held constant, F slightly outperforms Fb (78–74%), but if one only reports the composite of the nonoverlapping studies and same-groups analyses it now appears as if Fb outperforms F (74–69%). This meta-analysis and these scales are used for illustrative purposes, and sometimes the contrasts between same-groups analyses and nonoverlapping studies are far larger than the results obtained here.

Readers might be surprised by how often meta-analyses do not show complete overlap across studies on tests or indicators, and in many cases the overlap may be limited or minimal. For example, in the Vickery et al. (2001) meta-analysis, which rank orders methods, there is minimal overlap in studies across a number of the indicators. Thus, it is difficult to discern the extent to which differences are a product of true contrasts in efficacy vs. inconsistencies in the magnitude of the EGP. If, as we think is the case, the EGP frequently accounts for more variance than any other factor, and if the magnitude of the EGP may differ markedly across studies of various indicators or tests, then *rank orders may be far off and even negatively correlated with true values*.

The inflated accuracy rates the EGP produces are likely to lead in turn to overconfidence in malingering assessment methods. In Chapter 1 (see the section on "Overconfidence"), we detailed the multiple adverse consequences that can result from inflated confidence, such as reduced accuracy and an increased tendency to make overly risky and harmful decisions. Furthermore, to the extent the EGP distorts the relative ranking of tests or procedures, it can

easily lead to nonoptimal or poor selection of methods. For example, a test with a true error rate of 40% might be selected over one with a rate of 25% because inequities in the background studies may make the first test appear to be more accurate than the second. Although one might suppose that an overly inclusive approach to battery construction offsets such possibilities, one obviously cannot include everything. Additionally, as set forth in the lengthy materials on data integration, the inclusion of relatively weaker variables, even if they are valid, commonly has a negative influence on decision accuracy. As also follows from inflated study outcomes, error rates in clinical and forensic application can be considerably greater than research suggests, and the proportion of false-negative and false-positive errors can shift dramatically. As we will show, a marked increase in false-positive errors may be a common outcome. Surely it would be good policy in meta-analyses to separate results for studies that do and do not use comparable groups and determine if the findings vary.

We do not mean to paint a gloomy picture. Although we believe the EGP is an extremely important and under-recognized methodological problem, we also believe that methods can be developed to measure, account for, and attenuate or correct its influences and that this should be among the highest priorities for researchers in this area. At the end of this section we suggest a number of possible research strategies and corrective approaches, and interested readers can also consult Ahern (2010) for further details about the EGP in general and potential strategies for addressing it.

The EGP stems from four basic sources, which apply to a wide range of research in psychology (and likely other areas of "soft" science). First, as already described, there is a sensible desire to form groups with as few false members as possible. A second source relates to the nature of many entities and constructs in psychology that merit our attention and interest. These entities and constructs are often highly heterogeneous or show wide variation within classes. Examples include "executive functions," "attention," "aphasia," and, of course, "malingering." Such heterogeneity frequently results in part from limits in our classification systems and knowledge, but it is also intrinsic to many of the classes and constructs with which we have to deal. There is, for example, probably less variation between bicycle spokes or water molecules than there is between malingerers.

A third source is the common lack of procedures that, in a sizeable percentage of cases, achieve definitive or near definitive accuracy in determining who does and does not fall within the class of interest or whether a complex of behaviors do or do not belong within the class. It is this difficulty or ambiguity that often motivates the researcher in the first place. As scientific knowledge advances, the percentage of individuals or occurrences that can be classified more definitively often increases, which redirects attention to the more ambiguous cases, creating an unsettling but imperfect paradox. We wish to learn how to classify the remaining ambiguous cases accurately, but in order to do so our research efforts are hampered by the very same problem we are setting out to address. To use a concrete, albeit inexact, analogy to illustrate the point, how could we study the nature and characteristics of trees if we often did not know how to identify trees in the first place? The dubious approach that may commonly be taken is to study the cases we do know how to identify in order to try to learn about the cases we do not know how to identify, which often produces pseudoknowledge that nevertheless fools us into thinking we are getting somewhere. We are better off realizing that the seeming paradox is incomplete, permitting means to attack the problem (most of which involve some form of bootstrapping, as we will describe).

Fourth, and related to the first three problems, the types of cases we do know how to identify are often extreme and unusually clean manifestations of the entity or construct under study and hence may apply minimally to the cases of greatest clinical interest and challenge. For example, when individuals perform below chance level on almost every test with a forced-choice format, the identification of malingering may approach 100% accuracy. However, including these individuals

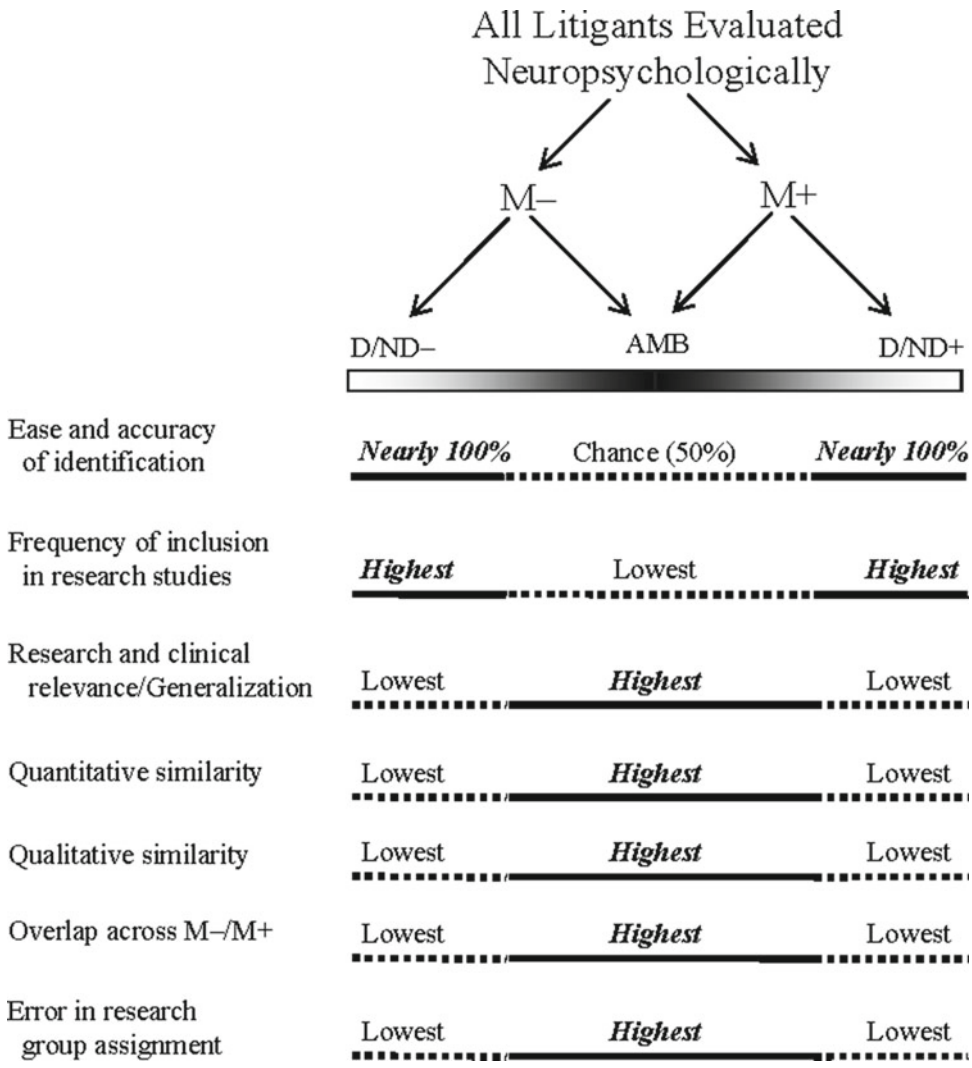

Fig. 2.6 Summary of extreme group problem (EGP) in relation to research and practice

in research groups in order to try to learn something about malingerers who are considerably more skilled and subtle may produce outcomes that are systematically misleading and do not improve but rather diminish accuracy, even to levels below that of a coin toss.

Figure 2.6 provides a schematic summary of the EGP and aspects of its interface with research and practice. For the sake of clarity, we start with the basic distinction between the presence and absence of malingering, and will later move on to the higher levels of complexity that are more typical in legal evaluations. Among all litigants for whom neuropsychological evaluation is performed, some are not malingering (M–) and some are malingering (M+). The certainty with which M– and M+ cases can be identified varies, and as before we have used the terminology D/ND (for definitive, near definitive) and AMB (for ambiguous) to reflect surety of identification. In Fig. 2.6, however, rather than treating these designations as dichotomous categories, we have placed them on a continuum. The D/ND M– and M+ cases occupy the extremes, and as one moves toward the middle from either end the cases become more ambiguous, with the most ambiguous

cases occupying the middle area. As we have emphasized, a major priority for research is reducing the percentage of remaining ambiguous cases. *Ease and accuracy of identification* represents another way of expressing standing on this continuum of definitiveness/ambiguity.

Examining the remaining entries and their relation to research priorities raises obvious concerns. Individuals who fall near the extremes of the continuum (e.g., definitive and near definitive cases, subjects in simulation studies) have the highest *Frequency of inclusion in research studies* on malingering. However, the cases of greatest *Research and clinical relevance/Generalization* fall in the middle of the continuum. For reasons we have touched on and will further elaborate shortly, studies on extreme cases may not only generalize poorly to more ambiguous cases but may well lead to reliance on indicators that minimally enhance or even diminish accuracy. *Quantitative similarity* and *Qualitative similarity* reflect the potential for changes on these dimensions when one moves from more extreme to less extreme cases. For example, failure on certain test items that research suggests is indicative of malingering may instead be more highly associated with true injury. (These sorts of reversals are not as unusual as they might seem, the literature on "scatter" and neuropsychological status providing multiple potential examples [see Faust et al. 2011].) *Overlap across M–/M+* also addresses the potential for qualitative shifts.

Error in research group assignment goes straight to the dilemma that ultimately creates extreme groups. When research groups are formed, all else being equal, error in group assignment is highly undesirable. For example, it would be exceedingly problematic if we drew research subjects from the very middle of the continuum, in which case about half of the subjects in the "malingering" group would be nonmalingerers and about half of those in the "control" group would be malingerers – we obviously would be pushing further and further into ignorance and error. We do not question whether erroneous group assignment should be a serious concern or whether, when it goes too far, it may not only inhibit progress but lead us in reverse (because much of what we thought we were learning would be wrong). However, minimizing error in group assignment should not necessarily trump all other considerations in research design, and if taken too far, as we believe too often happens in malingering research, it may greatly inhibit research progress. In part, whether we are going too far can be measured by examining the extent of the EPG. In many cases in science, using valid but fallible indicators for group assignment and then applying, to the extent possible, means to account for group impurity, is a crucial or even necessary step for achieving progress. Fallible group assignment is *not* a desirable end point, but it can be a necessary means for moving in the right direction.

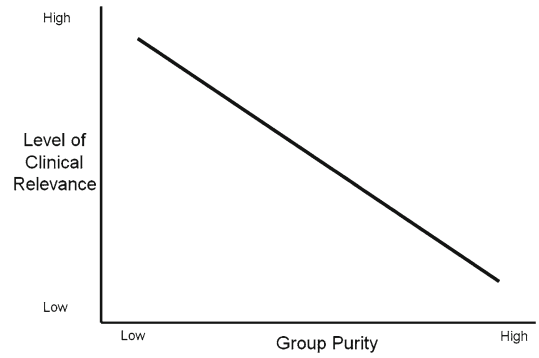

Fig. 2.7 Relationship between research group purity and degree of clinical relevance of research findings

Figure 2.7 illustrates the investigator's dilemma in malingering research. Put simply, in the usual circumstance, the purer the group the less relevant or helpful the research in addressing the most pressing current clinical needs (and the greater the likelihood the findings will be misleading). Across various areas of science, such associations tend to hold to the extent the four factors or sources that create the EPG are present. In malingering research, at present, group purity almost always comes at the cost of selecting extreme cases. Thus, where clinical needs are the greatest, overprioritizing purity renders too much research of limited value or even potentially misleading. It is essential to increase the level of clinical relevance without introducing too much error in classification or to find means to compensate or account for it.

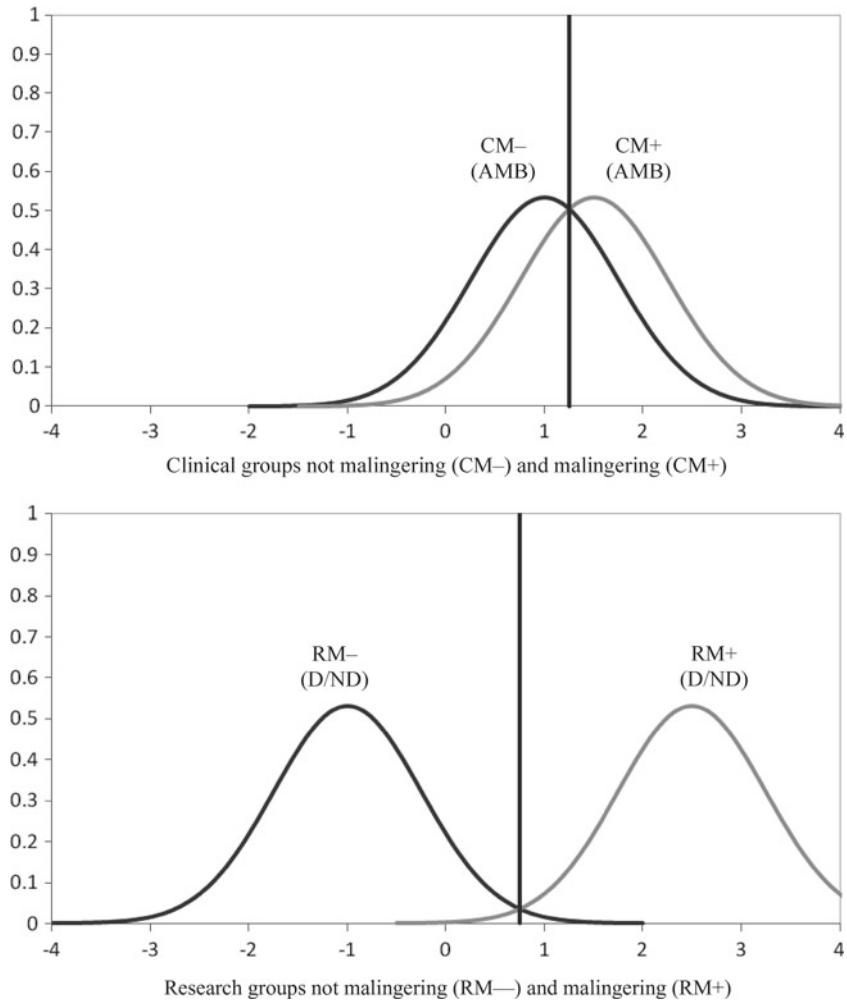

Fig. 2.8 Differences in score distributions for clinical (more ambiguous) cases and research (more definitively identified) groups

The potential for quantitative and qualitative shifts and the resultant impact on error rates can be explained and illustrated through a series of figures. Staring with quantitative shifts, we turn to Fig. 2.8. To distinguish between malingerers and nonmalingerers in applied vs. research settings, we will use the designation C for clinical and R for research. We know there can be overlap in these groups. For example, a researcher may cull presumed malingerers from applied settings, and the intent here is to first demarcate the totality of individuals in applied settings vs. those that make up research groups.

The upper graph in Fig. 2.8 depicts ambiguous cases, those that tend to present the greatest diagnostic challenges and hence create the greatest research needs. In many such cases, there is a viable basis to suspect malingering. Of those suspected of malingering, some are malingering but others are not, and distinguishing between the two is not straightforward. We have previously referred to these subtler or less extreme cases with either the designation MS (suspected malingerers) or AMB (ambiguous cases), the latter of which appears in the current figure. In the top of Fig. 2.8, CM– (AMB) designates individuals in

clinical or applied settings who are not malingering and CM+ (AMB) individuals who are malingering, with the AMB added to both groups to indicate that the cases are not obvious.

In the top half of Fig. 2.8, we have drawn hypothetical distributions that might apply to a modestly robust indicator, such as a score on a malingering subtest that achieves a .50 separation between the groups. Note that *both* groups have elevated averages relative to the mean of zero for the indicator, something which is often to be expected because individuals are suspected of malingering for a reason. Thus, it is not only the group that is *suspected of malingering and is* malingering that achieves elevated scores, but also the group that is *suspected of malingering but is not* malingering. What distinguishes the groups is not elevated scores per se but relative levels of elevation. In Fig. 2.8, the CM− group has a mean score of +1.00 SD and the CM+ group a mean of +1.50 SD. As one can see by consulting research on various tests, such as the MMPI-2 (see Greene, 2011), it is quite common for individuals with genuine disorder to score above the mean on malingering indices, one reason being that detection rests in part on overendorsement of items that have a true association with pathology. Consequently, in comparison to normal groups, scores are deviant, although usually only to a modest degree. Also, partly because so many effect sizes in malingering research are grossly inflated, one might think a figure like .50 is feeble or unusually low. However, .50 reflects a fairly robust and helpful relationship, especially for an isolated indicator or score. With these means for the two groups, the optimal cut would likely fall at around 1.25 (although one might want to shift it to the left or right if false-positive or false-negative errors were a higher priority).

The bottom graph in Fig. 2.8 depicts distributions on the same indicator for research groups whose members have been identified definitively or nearly definitively (D/ND) as not malingering (RM−) or as malingering (RM+). Based on having examined many malingering studies, we would submit that the distributions we have drawn for these two groups are not rare. As one can see by referring back to Fig. 2.5, which was based on the Rogers et al. (2003) and Vickery et al. (2001) meta-analyses, between 22 and 32% of the outcomes reached or exceeded effect sizes of 3.00. When one examines research across a range of malingering indicators, it is not difficult to find extraordinarily large effect sizes. Naturally enough, if authors point to indicators that yield the highest effect sizes as the most valuable ones, then practitioners will often be operating on the basis of distributions much like the ones that appear in the lower part of Fig. 2.8 (even though they are mainly an artifact of the EGP). We have not tried to represent the worst case scenario or something even close to it, such as the third and fourth columns of effect sizes listed earlier (i.e., to save the reader from backtracking: 5.30, 5.47, 5.74, 6.53, 8.14, 10.24, 10.38, and 13.66).

Compared to the CM+ group (the clinical malingerers) in the upper portion of Fig. 2.8, the RM+ group has shifted 1.00 SD to the right (from +1.50 SD to +2.50 SD). In comparison to the CM− group, the RM− group has shifted 2.00 SD to the left (from +1.00 SD to −1.00 SD). Each research group is more extreme than its corresponding clinical group, although the shifts are unequal, with the CM− group shifting more than the CM+ group, hence changing the optimal cutting score from about +1.25 to about +0.75 SD. Asymmetrical shifts are probably common, with the CM− group changing more than the CM+ group. The tendency toward asymmetrical shifts can be explained as follows. In clinical settings, we are starting with groups that are suspected of malingering (CM− and CM+), both of which usually obtain above average or elevated scores on malingering indicators. Consequently, in research settings, when we seek a group for which malingering is a near certainty (RM+) and another for which nonmalingering is a near certainty (RM−), we often need to move further from the clinical baseline in the latter instance because the standing of the CM− group is likely to fall toward or in the abnormal range. The typical subject for whom there is no reason to suspect malingering is often "cleaner" than the typical clinical case or even control case, and thus the distance one must travel along the distribution of scores to reach something approaching definitely

not malingering is often past the point of normality (relative to scores on the measure of interest).

The impact of these shifts, especially when asymmetrical, can be extremely destructive. First, research studies will produce highly inflated accuracy rates or effect sizes, which is exactly what happens on many occasions. The worse the EGP, the better the method will look. Second, to the extent the magnitude of the EGP varies across studies examining different malingering indicators and tests, the greater the distortion in the relative efficacy of methods. We think there are strong reasons to posit that the EGP often exerts a far greater impact on study outcomes than the intrinsic properties of measures, and hence rank orderings of procedures are highly error prone, leading practitioners to frequently substitute weaker methods for stronger ones. Third, and perhaps most concerning, asymmetrical shifts distort optimal cutting scores. If the shifts are greater for the control group, which we think is the more common occurrence, it will increase the false-positive rate, and if the shift is in the other direction it will increase the false-negative rate. One can see in Fig. 2.8 that the optimal cutting score has shifted .50 SD to the left, the result being that the CM– group mean exceeds that cutting score by about .25 SD. As a consequence, *about 60% of those who are not malingering will be misidentified – the false-positive error rate is now greater than the results obtained by flipping a coin*! In a criminal context, should the shift go in the other direction, violent offenders feigning mental incompetence might be missed in a large percentage of cases. Although these graphs are hypothetical, the basic phenomena described here are real, and a magnitude of error that equals or exceeds that set forth in this example can be expected at times.

Thus far we have illustrated what we call quantitative shifts. Qualitative shifts can also occur and compound error. We do not wish to enter into pseudodebates about qualitative and quantitative indicators because, as noted in Chapter 1, almost any qualitative indicator can be quantified, rendering many arguments about relative merit moot. Suppose instead we make the distinction between continuous variables and dichotomous variables, the latter of which covers almost all the forms of "qualitative" data referred to in these debates. We would simply argue that all forms of data should be subject to formal study and testing when possible and judged on the basis of scientific merit rather than ideological positions. Surely at times, dichotomous distinctions can be of value (e.g., breathing or not, bizarre delusions present or absent, operable or inoperable tumor, performance is or is not well below chance on multiple forced-choice methods). Various characteristics or red flags have been proposed as malingering indicators, a number of which can be conceptualized as dichotomous and which may well have value.

With common study designs, investigators recruit research group members they can identify definitively or almost definitively to try to learn how to identify clinical or forensic cases we do not know how to identify. This approach more or less guarantees some differences at the outset between the study subjects and the ultimate group to whom we wish to generalize the research (the AMB groups). If those in the AMB groups shared the characteristics of the individuals we can identify as D/ND malingerers and nonmalingerers, then they would not be AMB cases. Additionally, because of positive and negative manifold in psychology (good things are usually associated with other good things, and bad things with other bad things; see Meehl, 1990), ineffectual malingerers that are relatively easy to identify probably differ in more ways than the indicators used to identify them for purposes of the study (e.g., they may be less intelligent on average, more likely to present highly implausible symptoms, feign too broadly and grossly, have more difficulty keeping track of lies, or make less effort to prepare). Similarly, the control subjects, who are usually individuals for whom there is little or no reason to suspect malingering, likely also differ from their counterparts.

Figures 2.9 and 2.10 illustrate what we refer to as qualitative shifts and potential reversals. For purposes of illustration, let us designate our qualitative indicator as sign X. Perhaps it is recognition memory markedly below spontaneous recall, early failures on easy test items, avoidance of eye contact, long response latencies,

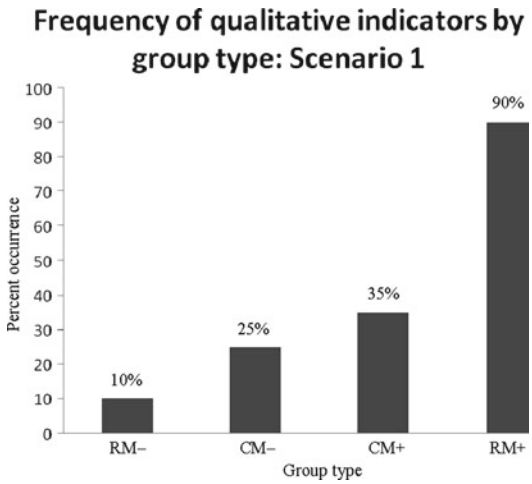

Fig. 2.9 Reduction in differential frequency from research to clinical groups

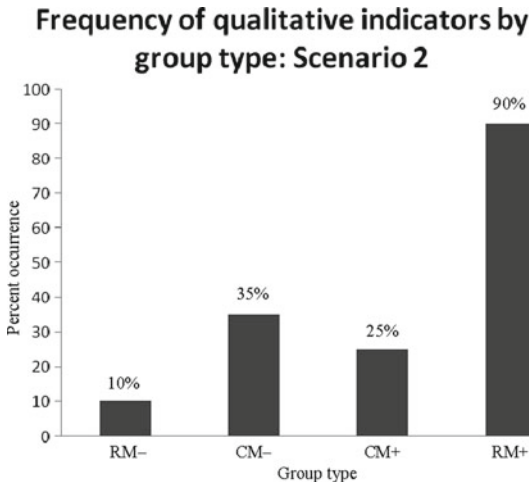

Fig. 2.10 Reversal in differential frequency from research to clinical groups

approximate answers, or some other such potentially differentiating feature. We may also have background studies (using extreme groups) that seem to support sign X. In any case, when selecting individuals for our current study, the presence of sign X is considered an aid for identifying malingerers. If it presents along with certain other potential indicators (e.g., failure on forced-choice items), that individual is selected for the malingering group. When examining exclusionary criteria for the control group, the presence of more than one sign of malingering eliminates the individual from consideration. Given these selection criteria, few individuals in the nonmalingering group demonstrate sign X and most in the malingering group do demonstrate sign X. As one can see in Fig. 2.9, the RM− group demonstrates a frequency of 10% and in the RM+ group a frequency of 90%. If, as is sometimes done, the composition of each group is itself considered instructive about the characteristics of malingering, then in rather circular fashion one might conclude that sign X is a strong differentiating sign, occurring 9 times more often in malingerers than nonmalingerers. (This, unfortunately, is almost exactly the type of circular process used when depending on clinical experience to try to determine the characteristics of malingering.) Additionally, almost anything that correlated strongly with sign X, say, sign Y, might also show similar differential frequency.

Given the strategy of group formation, however, the relative frequencies of sign X or sign Y might be very different in applied settings. In Fig. 2.9, sign X is still valid for distinguishing the clinical groups (CM− and CM+) but not nearly as strong an indicator as the research study suggests, and, depending on base rates and the availability of other indicators, a practitioner might be better off not using it at all. For example, if a strong alternative indicator is available, it will often conflict with sign X, and in the great majority of instances when one defers to sign X over the stronger indicator it will result in error.

Figure 2.10 illustrates what we refer to as reversal. Here, the relative frequency of the qualitative indicator is reversed among individuals in the applied setting; that is, the presence of the sign is in fact more common in the nonmalingerers than the malingerers. Lack of validity or reversal is not an outlandish outcome when extreme groups are selected because the characteristics of these groups and correlated features are unlikely to generalize to the AMB cases in applied settings. For example, gross failure on malingering may be common among research subjects but almost never observed among AMB cases. It is not hard to generate potential situations in which reversal might occur. For example, more severe cases of

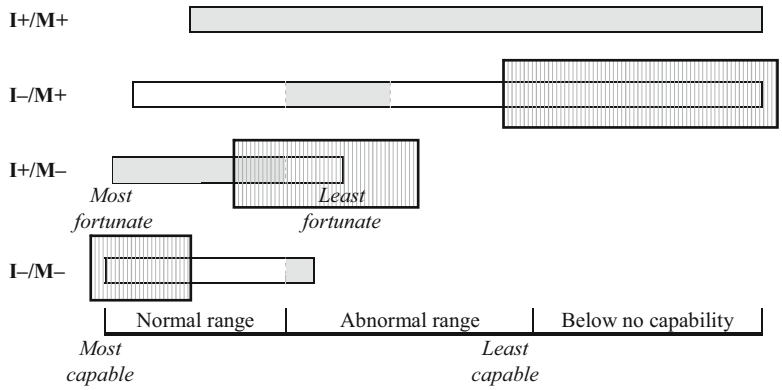

Fig. 2.11 Relation of areas of concern to research groups (patterned boxes)

PTSD may be associated with higher rates of noncompliance. More severe brain injury cases may be associated with considerable response latencies; greater inconsistencies in performance due to such factors as easy fatigability, increased impulsivity, and attentional lapses; or certain elevations on personality tests that could be mistaken for overreporting or antisocial tendencies.

To summarize to this point, potential differences in amount and kind are often major obstacles to generalization from research studies to applied settings. They not only may distort or inflate accuracy rates, sometimes leading to gross overestimates and dangerously inflated confidence, but they often do so unequally across studies and analyses of indicators, obscuring or reversing their relative standing. Furthermore, asymmetrical shifts in the extremity of malingering and control groups from research to clinical settings may alter cutting scores, markedly increase the frequency of false-negative or false-positive errors, and lead us to believe that qualitative indicators that are minimally effective or that even reverse indicators of group status have considerable value. In many instances, much or all of this may be happening under our noses without our recognizing what is occurring. To those who take this to mean we are better off if we instead trust our clinical judgment and experience, it is highly likely that each and every one of the aforementioned problems will be no better and likely worse should those alternatives be selected. We will not work our way out of these problems experientially, but only by well-directed scientific efforts. Almost all of the considerable gains in malingering detection have ultimately been achieved through research (which may use clinically-based observations, conjectures, or insights as crucial starting points), and there seems to be no compelling reason to think this situation will change. However, the distance remaining to be traveled in malingering detection may be a good deal further than is sometimes assumed.

It is sobering to think that we have set forth a simplified set of circumstances. Figure 2.11 (an elaboration on Fig. 2.1) displays relationships between the EGP and joint presentations. For each of the four groups, the respective shaded areas of concern reflect overlap with other groups and hence ranges of outcomes that often create the greatest diagnostic difficulties. For example, if an individual who is not injured but is malingering obtains extremely poor results on forced-choice testing, the presence of malingering is likely to be recognized (although, arguably, a false identification of injury could still occur). The interested reader can turn back to the earlier section, "Mixed Presentations: Injured and Malingering," which covered the rationale for the positioning and widths of the shaded areas, and there is no need to reiterate those points here.

The boundaries set forth in the patterned boxes that are superimposed on certain sections of the entries for the I–/M+, the I+/M–, and the I–/M– groups identify typical compositions of research groups. For example, when the intended study

group is malingerers (who presumably are not injured as well), researchers often focus on extreme cases to minimize error in group assignment; similarly, clean or extreme cases may be selected for a normal group. For injured groups, researchers often seek to identify individuals who clearly are injured and clearly are not malingering. The end result is that very little research focuses on AMB cases for which the diagnostic challenges are greatest; furthermore, due to potential quantitative and qualitative shifts, what is learned might not get us very far or may even be frankly misleading. Finally, the I+/M+ group contains no box identifying typical compositions in research studies because, despite the great importance of this category, there is almost no research on it. If the reader finds himself or herself getting a methodological stomachache at about this time, we can only say that all of the authors have shared the feeling. However, no one ever said that good science was easy, and we believe that these problems can be addressed productively through concentrated effort. A number of suggestions follow.

Possible Strategies for Addressing the EGP

It is sensible to be concerned about error in group assignment, but not to the point of generating research so encumbered by the EGP that it is of little or no value or even systematically misleading. Although minor or even modest problems in this area might not be so damning for exploratory projects in the context of discovery, it is a major shortcoming in the context of verification. Research on two basic fronts may assist in attacking the problem. First, recognition, measurement, and attempts at attenuation or correction are all worthy goals. Second, rather than learning to live with the problem or devising means to lessen its influence, we would be better off avoiding it in the first place. We will address both areas here, and more detailed discussion can be found in Ahern (2010), Bridges et al. (2009), and Faust et al. (2009b).

There are various ways to identify and measure the EGP. Examining the formation of research groups is one key. For example, cues to the presence and extent of the EGP include the number and breadth of inclusionary and exclusionary criteria and the percentage of potential subjects eliminated from a study. Another tip-off is wildly fluctuating accuracy rates or effect sizes across studies on the same measure. One can examine whether "accuracy" seems to vary systematically with the extremity of groups, and how closely those groups resemble the cases of clinical interest. Large or outlandish effect sizes are strong indicators, as are implausible accuracy rates.

In some cases, accuracy rates exceed the level possible given limits in the reliability of measures, this occurring because one is not studying a representative sample of cases but rather cases toward the extreme ends of distributions. Reliability figures reflect not only the intrinsic quality of tests but also the extremity of the groups studied. Thus, for example, depending on the metric used, an analysis of reliability based on a broad distribution of cases can yield a lower result than examination of extreme cases drawn from the far ends of the distribution. Suppose, for example, as is sometimes done, the consistency of classification is taken as an indicator of reliability. Here, if one mainly draws cases with very high or very low test scores, then even if there is considerable variation in results on retesting, decision consistency can still be very high. By way of analogy, if we are examining the consistency of first base umpires' decisions but primarily limit ourselves to cases in which the runner is either out or safe by a wide margin, then even if decision consistency is substandard on many calls and more typical situations, very high consistency rates may still be obtained. If a method with reliabilities in the .60s or .50s when used with broad samples generates accuracy rates in the 90% range in a separate study, there is a very good chance that the EGP is operating.

It is almost always worth checking within or across meta-analyses that compare the efficacy of different measures or indicators for the same

diagnostic categories or outcomes. When study groups overlap entirely, the EGP is held constant, and this should often reduce or eliminate its confounding effects on the relative performance of measures. Of course, to the extent the EGP is present, accuracy rates may still be grossly inflated, cutting points might be shifted, and reductions in or reversals of qualitative indicators may still occur. Nevertheless, barring interaction effects, the relative merits and rank ordering of methods should be preserved. If some studies have group overlap and some do not, one can compare outcomes across the overlapping and nonoverlapping studies to look for trends. We would humbly suggest that journal editors keep this problem in mind and require that comparative analyses of methods separate the overlapping and nonoverlapping studies and examine whether systematic differences result. Test A might beat test B when study groups overlap, but a meta-analysis may have pooled the overlapping and nonoverlapping studies and altered the comparative outcome. It may be a mistake to think that the presence or absence of overlap is unlikely to be systematically related to the performance of the same indicator (e.g., the F Scale on the MMPI-2). For example, better designed and well-funded studies may be more likely to include a broader range of indicators.

Probably the best solution to the EGP is to recruit representative samples. This is often a difficult undertaking for a number of reasons, in particular because one would need accurate methods to identify positive and negative cases across the range of possible (or at least relatively frequent) presentations, and it is the need for such knowledge that often drives the study in the first place. If we had this knowledge we probably would not need to perform the study, and we are undertaking the investigation because we lack this very knowledge. Approaches that may assist in recruiting more representative samples (e.g., Group Membership by Chance) are discussed below.

In the meantime or as supplemental strategies, researchers may feel freer to recruit more relevant but less pure groups if approaches can be used to assess or adjust for error in group classification. One such approach is mixed group validation, also described below. For the remainder of this section, we wish to lay out what we have elsewhere labeled the *Definitive/Near Definitive Variation Rate* (DVR). We first described this method in Faust et al. (2009b), but it is possible we are unaware of precedents from which we have unwittingly borrowed and hope we are not failing to properly credit the work of others.

As we have discussed, restricting studies to extreme cases is likely to yield misleading results. A fundamental factor impeding generalization is that the procedure used to form research groups depends on ease of detectability or characteristics that set these participants apart from those that are not selected. However, it is the latter group – the group we presently cannot detect or have greater trouble detecting – that we are trying to determine how to detect more effectively. Group formation is thereby inevitably tied to a feature (detectability) that distinguishes the research subjects from the group we want to learn about, and that feature may also be associated with various other characteristics that also separate these groups. The ultimate result is often lack of generalization to the group of greatest clinical interest or, even worse, "indicators" that are negatively associated with malingering (reversals).

The DVR strategy capitalizes on the occurrence of D/ND cases. Assume that across groups of malingerers (M+), the percentage that can be identified definitively or nearly definitively (D/ND M+) is fairly constant, and that the D/ND rate is also fairly constant for nonmalingerers (M−). Although the D/ND M+ rate and the D/ND M− rate each need to be fairly constant, the respective rates do not need to be consistent with one another. For example, it would not matter if the rate for the D/ND M+ cases is twice as high as the rate for the D/ND M− cases. Furthermore, when starting out, one does not even need to know what either of these rates might be, so long as there are strong reasons to assume they are both significantly above 0%, which is certainly the case. For purposes of illustration, we will assume a hypothetical D/ND rate of 40% for both the M+ and M− cases (leaving 60% from each group as ambiguous cases).

Assuming that 40% of malingerers can be identified definitively or nearly definitively, it also follows that *if* we could randomly select, say, 1,000 malingerers, then about 40%, or 400 subjects, would be classified as D/ND M+ cases. (We use the qualifier *if* because at present there is no method to identify such representative samples of malingerers.) Conversely, *if* we were able to randomly select a representative sample of 1,000 individuals who were *not* malingering and evaluated each one, 0% (or close to 0%, given the potential for some error) would be classified as D/ND M+ cases. We have also assumed that about 40% of nonmalingerers can be identified definitively or nearly definitively. Consequently, working from our hypothetical sample of 1,000 malingerers, 0% (or close to 0%, given the potential for some error) would be classified as D/ND M− cases. Among the hypothetical sample of 1,000 nonmalingerers, about 40%, or 400, would be classified as D/ND M− cases.

Although we are about to add one more set of hypothetical figures, we wish to emphasize that all of the figures set forth in this section are being used solely for illustrative purposes. Use of the DVR procedure also does not require, as noted, knowledge of the D/ND rates for malingering or nonmalingering groups, nor does one require knowledge of base rates for malingering. Furthermore, it is not necessary to identify representative groups of malingerers and nonmalingerers. Even evaluations for the occurrence of malingering and the separation of individuals into D/ND vs. ambiguous cases do not need to achieve a high degree of accuracy. The more accurate the classifications the better, but the procedure should be able to tolerate even a moderately high error rate. The critical point for now is that the total number of D/ND M+ cases and D/ND M− cases should vary markedly (in this illustration from 400 to about 0) across the extremes, that is, depending on whether one is drawing from a sample with all malingerers vs. a sample with no malingerers.

Again, working with a hypothetical figure, assume that the base rate for malingering among litigants seen for neuropsychological evaluation is 15%. Given this base rate, if one draws a random sample of 1,000 such litigants, 150 individuals will be malingering and 850 will not be malingering. (We realize by dichotomizing the presence or absence of malingering we are simplifying matters and disregarding joint presentations, but again our major intent here is clarity, and the same principles should apply with more complex situations.) If the percentage of individuals that can be identified as D/ND M+ is a relative constant and falls at about 40%, as we have assumed for this illustration, then 40% of these 150 malingerers, or about 60, will be so identified. If the percentage of individuals that can be identified as D/ND M− is also about 40%, then about 340 of the 850 nonmalingerers will be so identified.

These potential outcomes can be summarized as follows, in each case assuming a sample size of 1,000:

	D/ND M+	D/ND M−
Condition 1. An all malingering group yields:	400	0
Condition 2. An all nonmalingering group yields:	0	400
Condition 3. Random sampling yields:	60	340

These are the identical outcomes that would result were the first group formed by using a variable with perfect accuracy in identifying the presence and absence of malingering and only positive D/ND cases were selected; if the second group were formed using this same variable and only negative D/ND cases were selected; and if the third group were formed using a variable with no validity (and consequently equated to random selection). We will use the term *comparison ratio* to refer to the result produced by a variable with no validity.

Suppose now we were able to draw a random sample of litigants undergoing neuropsychological evaluation, thereby providing the needed comparison ratio for research with this group. Based on background knowledge, we have estimated the base rate for malingering in our sample as modest to relatively low (e.g., 15%). Hence, we have a good idea about outcome if we identify all of the D/ND M+ and D/ND M− cases in our sample: we will have a considerably lower

number of M+ as opposed to M− cases, with a ratio approximating the one that appears under Condition 3 above, which is about 60:340, or about 1:6 (rounding off to the nearest whole number) As noted, this is the same ratio expected if a variable had no validity, which provides the foundation for its use as the comparison ratio. We need not know this ratio in advance; we derive it through random sampling of the overall group of interest, followed by evaluating the sample and identifying D/ND M+ and D/ND M− cases.

In contrast to a variable with no capacity to differentiate between group members, as validity increases, the comparison ratio will shift accordingly. Consider Condition 1, which illustrates the hypothetical result expected with a variable at the far end of the spectrum, or one with perfect accuracy in identifying the presence and absence of malingering. Here if we select the first 1,000 individuals with *positive* results, evaluate them, and identify D/ND M+ and D/ND M− cases, the obtained ratio should be about 400:0, which is far different than the comparison ratio of 1:6! If this same variable is used to select the first 1,000 persons with *negative* results, the obtained ratio should be about 0:400, again an extreme departure.

Although we would almost never anticipate such huge shifts, it does follow that the more valid a variable for separating group membership, the larger the shift. Therefore, it would seem feasible to measure a variable's validity and also to place it along an ordinal scale that reflects relative level of validity: the greater the shift, the higher its standing on the scale. One could also examine the impact of combining variables, such as the extent to which adding a new variable yields incremental validity.

The potential value of the DVR method is that it does not require knowledge of base rates or knowledge of whether individuals are or are not malingering for the group studied as a whole, and it likely can tolerate substantial departures from representative sampling. We realize we have only presented the broad outlines of this strategy, it is in an early stage of development, and considerable further refinement is needed. A number of practical obstacles would also need to be addressed. We would not expect such research to be undemanding but do believe that the DVR method is feasible. Given the scope and importance of malingering assessment, the effort and resources that would be needed to appropriately test and develop this method seem to be justified.

Lack of Representative Samples

Identifying representative samples of malingerers and nonmalingerers (and mixed presentations of malingering conjoined with disorder) would obviously be of great benefit. Representative samples are crucial for determining which features are valid predictors and differentiate among groups, appraising generalization of signs and indicators, and deriving accurate base rates. Unfortunately, researchers are often faced with one of two problematic situations. In one they have recruited a group whose members are known to be malingering with near certainty, but with an assemblage that is almost surely nonrepresentative of malingerers as a whole, especially the cases we currently have difficulty detecting and most need to identify. This is a variation of the EGP discussed above. In the other circumstance, a group has been identified that is known to be relevant, but within that group one does not know in many cases who is and is not malingering. The latter circumstance almost always holds in contrasting group designs. Thus, we may be able to obtain a group representative of those applying for disability, but we do not know who is malingering or to what degree, except perhaps for those who produce extreme outcomes and hence are not the cases we are trying to learn how to detect more effectively. The problem of determining the status of group members has limited the utility of contrasting group designs, although we believe there may be ways to augment these approaches to increase their effectiveness.

The seeming paradox is that one would need to know how to identify malingerers before recruiting representative samples, at which point one would not need to do the studies. The absence of representative samples, or rather the inability to determine whether samples are representative, greatly hinders efforts to identify and evaluate

potential malingering indicators. Under such conditions, it is very easy to inadvertently adopt signs that are ineffectual or, even worse, increase the number of misidentifications.

Situational variables may also separate research subjects and settings from litigants in applied situations. Many malingerers, especially in brain damage cases, have experienced an injurious or potentially injurious event. Thus, for example, a researcher might try to recruit subjects in emergency rooms who were in car accidents but did not suffer head injuries. Some malingerers have been exposed to models or mentors (e.g., a relative who has been injured or someone who has malingered successfully, such as a fellow prisoner). Numerous malingerers have met with attorneys before undergoing examinations, and a sincere attorney may provide inadvertent cues through leading questions about head injury, or may warn the client about tactics the independent examiner might employ. Many malingerers have also been subjected to multiple medical examinations, including those in which feedback or "education" about injury is provided. For example, a neuropsychologist who discusses results with examinees may provide detailed information about head injury or even about his reasons for questioning the examinee's cooperation. If the attorney is unhappy with initial assessment results, a new examination might be sought and the prior examination not disclosed, with the plaintiff now far better forearmed to influence outcomes in a desired direction.

We would like to propose an approach that we think offers promise for obtaining more representative groups of real-world malingerers. We label this the *Group Membership by Chance* (GMC) strategy, and we believe it can be applied to a range of situations in the social sciences when conventional methods of random selection are problematic either because of ethical constraints (e.g., head injury studies) or because means for identifying individuals with the condition in question are weak or lack adequate validation. In usual circumstances, in order to obtain representative samples, one selects randomly from a known population. Thus, were it feasible, one would randomly sample the population of malingerers and then compare that group with other groups the clinician needs to distinguish. Unfortunately, it is not currently feasible to do so and we are generally limited to samples that are almost surely not representative – and very possibly systematically misrepresentative – of malingerers as a whole. The more basic problem is the absence of a method for evaluating just how representative that subgroup might be. Without such a method, even if the researcher happens to obtain a representative group, one cannot determine that this good piece of fortune has occurred, and hence it really does no good.

Some malingerers are caught primarily because they are ineffectual malingerers. Others are caught primarily because they are unlucky. Take the following case in which one of the authors consulted. One of the professional staff, who had left the treatment setting at an unscheduled hour as a result of an unexpected personal circumstance, just happened to observe a patient, who momentarily let down his guard once he was blocks away from the hospital, exactly at the moment he engaged in an activity he absolutely should not have been able to do. Or a plaintiff may have just happened to run into an unusually skilled and determined detective who caught him acting normally, whereas seven other malingering coworkers happen to have been assigned to more mediocre sleuths.

In the idealized instance, an individual who is caught entirely as a result of bad luck is directly parallel to a malingerer drawn randomly from the pool of malingerers, that is, she represents in essence an instance of random selection. If one can identify enough such individuals, one should be able to comprise a group that is likely to be representative of malingerers as a whole, or at least a good approximation. This allows not only for the analysis of that GMC group, but also for checks on the representativeness of groups formed in other ways (e.g., malingerers caught by other means or cases compiled via contrasting group methods). It might also be possible to estimate the relative purity or base rates for malingering in contrasting groups, which offers major benefits, especially when studying generalization of measures across applied settings. For example,

using methods designed by Dawes and Meehl (1966), if one can determine the relative impurity of validation groups, one can then adjust for cases of false inclusion (i.e., the mixture of properly and improperly included individuals). The D/ND method described earlier would also benefit from informed estimations about the mix of group members.

There are a number of questions and issues one might raise about the GMC approach, some of which can be touched on here (see also Faust, 1997). One question involves the methods used for determining the level of chance in identification. We think that this is not too difficult a methodological problem because: (a) the method does not require perfect indicators (one does not have to be particularly concerned about some impurity), (b) rational analysis should provide reasonable accuracy in estimating the contribution of chance, (c) failures of inclusion (false-negative errors) do not have distorting influences (one can be conservative if need be without worrying too much about consequential problems with representativeness), (d) the approach described here is an initial approximation to addressing what has been a longstanding and very difficult problem and can be refined over time, and (e) a variety of checks can be built into the procedure. For example, a series of risky predictions can be made that should hold if the method works.

A second problem is not conceptual or methodological but practical. How could one possibly find enough caught-by-chance subjects? It is probably unnecessary to limit the method to pure cases because if the level of chance can be estimated even approximately and accounted for, more lenient inclusionary criteria would probably be workable. Nevertheless, data pooling would seem essential. On a national level, there are surely many such cases. The question is how to garner them. This is one of various domains of malingering research in which efforts would be helped greatly if more funding were available to researchers. Given the presumed cost of fraud attributable to malingering, these might be dollars well invested.

Base Rates: Some Research Priorities

In Chapter 1, we described the value and utility of base rate information. In a range of situations, base rates are among the most useful, or the single most useful, diagnostic indicator or sign. Additionally, knowledge of base rates is often critical in determining the potential utility of test results or other assessment methods. Shifts in base rates alter ratios between true-positive and false-positive, and true-negative and false-negative identifications. Base rate information is needed to determine whether we should use signs at all, the accuracy that signs achieve, and whether and how we should adjust cutting scores. As discussed in Chapter 1, professionals often seem to underweight base rates or have problems applying them properly, which can be viewed as a high priority item for education and training programs given the benefits accrued from better practices. Fortunately, an increasing number of publications in neuropsychology that address diagnostic practices in general or malingering more specifically, as well as professional manuals, touch on the importance and application of base rates. On occasion these discussions arguably conflict with sound advice by instructing individuals to formulate and apply composite base rates for practice settings, by underemphasizing the limited value of global base rates, and perhaps by overstating the surety of current base rate information (see "*Recognizing flawed advice about the use of base rates*" in Chapter 1). An even more fundamental problem is that articles or manuals may offer base rate estimates that vary widely and one generally does not know which estimate is most accurate or applicable in the setting of interest. Alternatively, global estimates may be provided, which are often of little utility.

The problem with global estimates is not only their occasional wide variance but the limited value of such information, in and of itself. First, these global estimates are mainly guesses, and although in some studies (but not others) practitioners show some congruence in estimates,

this is a soft evidentiary basis for determining accuracy. Second, the frequencies depend on a range of assumptions that may contain arbitrary elements, rest on insufficient knowledge, or do not address essential considerations. For example, frequencies will depend on where thresholds are set. If we equate almost any form of exaggeration with malingering, we are likely to obtain extraordinarily high rates, but if we set more stringent standards rates will likely decline sharply. This is a little like deciding the threshold for identifying friendliness and then claiming a certain resultant base rate is accurate. Third, almost none of the background studies include the conjoint category of malingering and injured, which, as we have described, can change obtained base rates and accuracy rates dramatically. A certain percentage of individuals who are malingering or exaggerating are also injured, and in some situations in which that frequency might be high proceeding to report a base rate for malingering (or a base rate for genuine injury) as if they were exclusive classes may be highly misleading. Fourth, and perhaps most importantly, when base rates for a condition vary widely across individuals, settings, and circumstances, which is almost surely the case here, global base rates are often of little utility. Such global base rates may minimally increase diagnostic and predictive accuracy, and in some cases may make no positive contribution or even diminish success.

We return here to the same type of paradox we encounter when attempting to determine how to best measure malingering – we need to know more than we know if we are to find out what we need to know. We need base rate information to appraise the accuracy of our diagnostic methods, and yet to determine base rates we need accurate measures of malingering. Nevertheless, it is commonplace in science to face such problems and yet to gradually evolve ways to overcome them, a process that is well underway in research on malingering detection.

Recognizing that global base rates are of minimal value, a key research priority is to determine how base rates vary across circumstances so that one can perform the type of reference group refinement described previously. The aim is to identify the base rate for the narrowest applicable group, with narrowness defined here by dimensions that: (a) alter the base rates and (b) are relevant to the individual under consideration. An obvious and important start is the presence or absence of financial incentive to malinger (e.g., involvement in legal proceedings), which, not surprisingly, seems to have a considerable impact on base rates (see Binder & Rohling, 1996; Frederick & Bowden, 2009; Reynolds, 1998). The larger the impact of variables, the less the remaining variance in base rates for which we need to account. Although one might suppose that a large number of factors are required, with these sorts of multivariate problems often a relatively small number of variables (perhaps three to five) are needed to reach or approach the ceiling in efficacy.

A number of investigatory strategies may assist in acquiring further base rate information and in identifying features that allow for determining differential frequencies among narrower groups. As already described, strategies for forming representative groups, such as the GMC method, may prove helpful. Meehl's taxometric methods provide a potential means for estimating base rates (see Meehl, 1995, 1999, 2001, 2004; Waller & Meehl, 1998). The *D/ND Variation Rate* (DVR) also may help in base rate studies. For example, in some circumstances, the percentage of D/ND cases within a sample may provide a strong cue for overall malingering rates within that sample (see further below).

A variation in contrasting group designs should assist in estimating base rates across situations and groups, with its utility enhanced if combined with the DVR method. One could develop a series of contrasting groups, each with likely differences in level of effort. It would be helpful to add groups with positive incentives to perform well, such as individuals applying for financial assistance for educational or vocational funding or individuals applying for employment. Other circumstances with positive incentives might include custody evaluations, certain types of competency examinations in which individuals want to perform well (e.g., competency to control one's finances or execute a will), and psychometric examinations

that are part of appraisals for resumption of driving privileges.

Some of these positive-incentive groups are likely to have malingering or poor effort rates that approach 0%, whereas the groups with the highest incentives to perform poorly may have rates that equal or exceed 50% or more. If, for example, one can determine or approximate the percentage of malingering cases that can be detected among all those who are malingering, and especially if this rate is reasonably constant (or at least predictable) across the groups, this should provide useful information about base rates. Suppose about 25% of cases can be detected with certainty or near certainty and this rate is relatively consistent across groups. One can then estimate the respective base rates for the different groups. In turn, the study of performance characteristics within and across groups might help in identifying valid and differentiating diagnostic signs and indicators and in identifying features that alter base rates and help in narrowing groups. If certain features appear much more commonly among the high frequency groups or show a steady rise in frequency as incentives or malingering rates increase, they are promising indicators or potential factors that alter base rates. A variety of approaches would likely be needed to advance or verify results, and in such bootstrapping operations one especially looks for convergence or consistency among different indicators as a key validation check (see Meehl, 1995). For example, it would be very interesting to examine whether potential indicators identified through such contrasting group designs were replicated in simulation studies. (In some circumstances, rather than starting with simulation studies and checking generalization to other circumstances, one could cross-check other research findings by subsequently performing simulations.)

Apart from contrasting group designs, there is at least one way researchers should be able to determine the lower limits of base rates. If one applies a measure with a very high true-positive rate, or measures on which positive results offer something close to *prima facie* evidence of malingering (at least on that task), then the obtained rate of positive identifications should provide a good estimate of minimum frequencies. For example, suppose we take performances that are well below chance on a forced-choice procedure as strong evidence for malingering. If this method was applied, say, to a group of disability applicants, the frequency of positive results should provide a minimal estimate of malingering rates. Of course, the true base rate might be substantially higher, but we would at least have a good approximation of the lower limit, and anything that allows us to start narrowing ranges is helpful. In many circumstances, even obtaining very rough estimates of upper and lower boundaries can give us clear pragmatic guidance. For example, some signs would prove effective, and some ineffective, anywhere within the range. Application of the strategy suggested here would probably uncover some situations in which our minimal estimates are erroneous, permitting us to sharpen our knowledge of base rates.

We might be able to do a good deal better in estimating minimal frequencies if we use multiple assessment devices or approaches with high true-positive rates, taking positive results of any of these measures as evidence of malingering. For example, we might look for positive results on symptom validity testing, direct evidence that the individual can perform normally in areas in which disability is claimed (e.g., videotapes), and instances of confession. Some individuals might confess at the time of evaluation, and others might confess if granted absolute assurances about immunity or after a nonreversible determination is reached. When formulating estimates in this manner, the conjunctive false-positive error rate of the measures would need to be taken into account. The major advantage of such a combined approach is reduction in the false-negative rate because, at present, approaches that appear to have high true-positive rates also seem to have high false-negative rates. We do not imagine that these types of combined approaches would be easy to pursue; but the effort would seem to be justified by the enormous benefits we gain if we are able to formulate reasonable estimates of the base rates. One would also think that the value of such knowledge should lead to favorable funding decisions.

Transparency

Most methods of malingering detection fall into one of four groups: they look for instances in which individuals perform (a) less well than they can, (b) less well than they should, or (c) differently than they ought to, or they (d) capitalize on stereotypic misconceptions about pathology (the last two categories could arguably be combined). These various approaches usually either depend on examinees holding some type of faulty belief, or they attempt to induce some false assumption. Attempts to induce false beliefs or assumptions vary in sophistication, power, and ease of detection (by the examinee). In some cases, an examinee is told that a test that is practically shouting out, "Try me, I'm easy," is really difficult, and then must perform miserably on the measure to be identified as a possible malingerer. In contrast, the MMPI-2 F Scale depends on false stereotypes about disorder, which may be shared by laypersons and mental health professionals alike (e.g., Gough, 1954). Simple attempts to educate oneself about disorder might not help. Rather, one needs to find out how the F Scale operates and how to identify F Scale items, and one then needs to endorse enough of those items to achieve an appropriate elevation but not so many that one is caught. For many methods of malingering detection, should the instructions or test stimuli fail to create misbelief or if examinees discern the simple, one dimensional detection strategy, there is a good chance the procedure can be beaten. And if the clinician interprets anything short of clearly malingered performance on one or a few such measures as presumptive evidence of good effort, the examinee is likely to beat the clinician as well. Many of our methods are much too transparent and are likely to lose effectiveness as word about how they work circulates.

It is prudent to assume that the underlying design of a malingering detection method will be discovered and circulated over time. The question is how to extend the time period before their efficacy is compromised or how to make them much more difficult to beat even if their underlying design is known. Given such realities as the exceptional motivation of some malingerers, the public nature of legal proceedings, the wide latitude given cross-examiners in challenging the underlying bases of conclusions, and the omnipresent Internet, it is unrealistic to believe that trade secrets will not leak out.

We can think of various means to counter transparency, and we are confident that others can expand and improve on the ideas provided here. First, problems with the transparency of forced-choice methods would be immediately improved by increasing the number of foils. Further gains would be realized by varying the number of foils across items and randomizing the order in which items with varying numbers of foils appear. For example, suppose one had items with two to four foils. Suppose that each of these items required the individual to identify a previously presented word on a memory test. Further suppose that the order of the two-, three-, and four-foil items was randomized, such that one did not complete the items with any particular number of foils in a group. The task of producing plausible yet varying rates of failure when trying to portray a serious memory disorder would seem to be far more difficult under such conditions than only needing to achieve a single believable failure rate. This and other approaches can capitalize on limits in human cognition, such as restrictions in the ability to track multiple dimensions of a problem simultaneously.

In a related vein, we might also take advantage of limits in human memory. For example, if inconsistency in presentation does help to differentiate between malingerers and the genuinely disordered, we can create circumstances in which fakers likely must resort to making up answers as they go along and will probably have extreme difficulty reproducing their performances at a later time. Suppose we compile a large number of items with low face validity that call for fairly rapid responses and have reasonable stability among honest reporters. A malingerer who does not know how she should answer but is trying to alter her presentation will most likely fall into an arbitrary pattern of responding that is very difficult to repeat on a subsequent occasion due to normal limits in recall. There are many other ways one

could attempt to design procedures that require extraordinary or impossible memory feats if one is to produce plausible performances over time.

Current attempts to create mental sets about item difficulty might be checked directly against subjects' perceptions. For example, how hard does an item seem at first blush and to what extent do suggestions about item difficulty alter perceptions (especially among those warned that the examiner may sometimes mislead them)? If we are going to pursue such approaches, we might try to expand and refine our methods for creating misperceptions. Indirect verbal suggestion might sometimes be at least as effective as direct suggestion (e.g., telling someone that they will get five chances at materials vs. telling them something is hard). Also, there would seem to be ways to alter perceptual impressions of difficulty without really changing objective difficulty, or in fact changing it in the opposing direction. For example, various perceptual illusions might be exploited to create misimpressions.

Other approaches might include shifting item pools and interspersing items that measure effort with items that measure ability. In the first instance, if, rather than having one set of items, there were numerous parallel items that could be used in varying combinations, it would probably make identification of the test more difficult and extend the half-life of methods. With interspersed items (which also might be combined with the first approach), one would have to be careful not to contaminate standard measures. Thus, it might be preferable to embed ability and effort items together in the development stage if, for example, one were simultaneously developing a new measure of immediate visual memory and ways of measuring cooperation on the measure. Although potentially complex, one major advantage is that in cases of poor performance, one would be able to simultaneously evaluate effort. Additionally, one could develop parallel forms with separate norms that exclude effort items in situations in which the appraisal of malingering is a low priority and one does not want to lengthen measures unnecessarily.

If and when baseline neuropsychological data become more widely available, formal approaches that calculate fit with expectations for preserved and diminished functions relative to the injury in question as opposed to inadequate effort might gain greater effectiveness and be relatively difficult to feign effectively. In the meantime, inter- and intraindividual variability and overreliance on subjective appraisal greatly curtail the efficacy of such approaches. In fact, they may be as much or more a guessing game for the neuropsychologist as the examinee, and the potential for error is a serious concern.

We may also want to explore methods that start with subjective ratings of item difficulty, probing for misimpressions. We might eventually be able to develop a fairly large set of items or be able to alter dimensions on the spot to examine their influence on impressions. Thus, rather than hoping or guessing that a misimpression has been created, one would wait to receive some confirmation that it has occurred before proceeding with the administration of items. We expect that in the future, much of neuropsychological and effort testing will use adaptive formats that make these sorts of procedures more readily achievable and routine.

With data pooling, it would be of interest to trace positive and negative rates on tests over time to provide a barometer of obsolescence. For example, if a measure that previously demonstrated, say, a 15% rate of positive results gradually showed a drop in comparable settings and circumstances, it would suggest that knowledge about how to beat the measure is becoming increasingly well known. More generally, when developing measures for assessing effort, examining transparency and vulnerability to knowledge of design could be considered essential and not something to pursue only after tests are published. These studies can include providing examinees with high quality information about the measure's detection strategies. Additionally, sequential testing might be conducted, in the first instance without information about detection strategies and then after feedback is provided about performance or results, as might be the case in legal settings when individuals are examined on multiple occasions by the same or different neuropsychologists. There is little doubt that in many instances an

individual who achieves results suggestive of poor effort is given some type of feedback about the outcome and at some later date is retested with the same or similar measures.

We are not suggesting that researchers abandon attempts to create measures that tap into false stereotypes. This approach has a long history of success, at least with the MMPI and MMPI-2, and we certainly should not demand from malingering detection devices that they catch everyone. Many individuals who malinger will not invest the time and effort needed to learn what they should do to effectively portray disorder, others will have difficulty mastering the needed knowledge and strategies, and initial evaluations may be performed before someone has a chance to become educated about the procedures. Methods that tap commonly held but false stereotypes may show limited redundancy with other approaches, which, as noted, increases their potential utility when combined with additional predictors. Furthermore, if methods and approaches are consistently updated as knowledge advances, it may be possible to stay a step ahead of many malingerers. For these reasons, efforts to extend this type of approach to structured and semistructured interview techniques and questionnaires that are specifically targeted at neuropsychological and related disorders seem very much worthwhile, as well as continuing efforts to study lay perceptions of head injury and other neurological disorders (e.g., Wong, Regennitter, & Barris, 1994). Such research can help in identifying candidate items for these types of malingering detection approaches.

Data Combination and Incremental Validity

As more malingering detection approaches are becoming available, showing that a single measure has discriminating power under one or another condition is minimally informative. Often, we will already have other measures that have passed the same basic appraisal, and one really needs to know how the new measure compares with other available devices and whether it makes a unique contribution to predictive accuracy. There is limited utility in identifying or developing indicators that are redundant with previously available methods. A study limited to showing that a new variable has discriminating power is usually of negligible help because we cannot evaluate whether that variable will have a negative, positive, or neutral effect on predictive accuracy when combined with other variables.

Rather, we should be trying to uncover variables that are likely to contribute unique predictive variance. It would be very beneficial if far greater effort was made to assess incremental validity, that is, any improvement gained by adding a new predictor to the best predictors that are already available. Given the inordinate demands that can be placed on subjective judgment, including the need to separate predictive and nonpredictive variables, gauge the strength of association between predictors and criterion, determine level of redundancy among predictors, and examine numerous possible ways of combining variables, it becomes imperative to draw on formal data combination methods, and particularly actuarial or statistical procedures. The development of the most effective decision methods, by its nature, requires study of incremental validity. Some investigators have examined multiple variables and their combined effects, which is a start; but too often these studies do not do much more than add to the innumerable demonstrations of a matter that is not at issue, that is, that the statistical combination of multiple valid predictors will usually outperform a single valid predictor. What these studies do not examine is the effect of combining new predictors with the best available predictors.

We cannot, however, perform all possible comparisons among measures and across conditions and variations of malingering. Blind empiricism is inefficient and usually ineffective in the long run. Rather, scientific efforts typically should be guided by principles, informed advice, and generalizations that usually hold. For example, it is completely impractical to test every conceivable comparison, and scientific and clinical activities often occur under conditions of uncertainty, in which there is no sure road. In such situations, however, one operating from

well-founded guesses and principles has a huge advantage over someone operating blindly. Take the methodological guide: "A method shown to make fine discriminations should do even better making more gross distinctions." There are times this generalization is flat-out wrong (as might occur, for example, when the situation changes qualitatively), but we usually do not know this in advance, and rather we are trying to resolve a question under ambiguous conditions. In attempting to do so, our odds of being correct are much greater if we follow this generalization than if we guess randomly, and correct guesses can greatly enhance the productivity of our scientific efforts. When designing studies, we should especially keep in mind the advantages gained by pooling nonredundant measures.

Caveats and Final Comments

Proposed Criteria for Malingering Detection

Various criteria for identifying malingering that investigators have proposed can facilitate research and communication within the field. However, these proposals are clearly experimental, and there are strong grounds to question whether they should be used at present in legal cases. Those proposing criteria are often open and explicit about their tentative standing. Nevertheless, some courtroom experts seem to act as if developers' cautionary statements should be disregarded.

To an extent that may not be recognized, definitions or criteria follow from scientific knowledge and advancement rather than the reverse. A questionable set of criteria based on insufficient knowledge can lead to nonproductive research efforts and misleading results. As an imperfect analogy but one that is illustrative, suppose we wished to learn about whales and prematurely defined them as animals that live in the sea, draw air from water, are all carnivorous, and are always over 50 ft long. Given such criteria, how much would we have learned, for example, about Belugas?

Similarly, operational definitions often do nothing to resolve critical conceptual issues, create a false sense of scientific resolution, and were abandoned by almost all philosophers of science decades ago (and nearly so by Bridgman (1927) at the end of his famous – or infamous – book in which he introduced the flawed concept). If malingering is what a malingering test measures, then that is what it had better do or the "definition" is erroneous. (However, repudiation of operational definitions should not be confused with the potential advantages of clear and explicit definitions, which is another matter.) Furthermore, as malingering is a hypothetical construct or latent entity, it follows that it cannot be reduced to a set of observations or observables because inference is always required. Thus, as the philosopher puts it, surplus meaning is involved, which should not be equated with a scientific sin.

Labeling

Labels assigned to results on malingering tests are sometimes highly misleading. Some test manuals or interpretive procedures adopt a very stringent threshold for identifying malingering, such as a probability of at least 90%. At times, even if the likelihood approaches this level but falls just short of it, the designated description or the expert's summary comment might be something like "within normal limits" or even "indicative of good effort." When a clinician indicates, for example, that an outcome is "unremarkable" or "confirms adequate effort," jurors probably would have no idea that chances might approach 9 out of 10 that effort on the test was poor or insufficient (but that a rather extreme standard was being used for the identification of malingering). Some experts describe almost any result that does not strongly indicate poor effort as demonstrating a satisfactory or high level of effort, which treats the matter as all or none and disregards degrees between these extremes. Similarly, defense experts who describe effort as inadequate or poor may not communicate how close the call was, or perhaps that the evidence was inconsistent. For example, they may emphasize one questionable result and underweight a number of other scores that fell within expected levels for the injury in question.

Descriptive or labeling practices should serve to provide accurate information and avoid misimpressions. Labeling practices often originate from the meritorious desire to avoid false-positive identification of someone as malingering (see further below), but to the degree labels create confusion in either direction, the trier of fact may well form inaccurate impressions. In general, it might be better to report both the probabilities and one's conclusions, instead of merely classifying the results one way or the other or providing an interpretation that is likely to cause misperceptions. Otherwise, the expert arguably is withholding critical information from the trier of fact.

Values Placed on Avoiding False-Positive vs. False-Negative Errors

An associated practice is to select cutoffs for tests that minimize false-positive classifications but lead to relatively high false-negative error rates. Again, the expert might select a very high threshold for identifying malingering, which produces few false-positive errors but possibly frequent false-negative errors. We question whether it is proper for *experts* to tip the balance *either way*, especially without disclosing these practices, because it potentially usurps the jurors' moral or decision-making responsibilities and surreptitiously substitutes the expert's personal values. (Arguably, the situation can be different in a clinical context, where there are often strong grounds to be very conservative about identifying malingering. In clinical settings, false-positive errors may cause considerably more harm than false-negative errors, and the moral obligation is to help the patient and, above all, cause no harm.) Once more, it might be best to report the outcomes of tests and procedures explicitly and then provide interpretations. Additionally, in forensic contexts, which error is worse is not necessarily obvious or can vary. Suppose a conservative interpretive strategy leads one to misidentify a criminal who plans to kill upon release as compliant with testing and as having a psychotic disorder. If this conclusion ultimately influences institutional transfer or release, should we necessarily view the intent to minimize false-positive errors at the cost of markedly inflating the false-negative error rate as prosocial or morally compelling?

Effort Is Not All or None

Overly general descriptions are sometimes too readily assigned to outcomes on effort tests. A person who does not exceed cutoffs designed to identify *poor* effort on a measure or two has not necessarily put forth *good* effort across the evaluative session. Alternatively, a brain-injured patient with limited endurance who is given a malingering measure a couple of hours into the testing session and obtains a depressed score may have exerted excellent effort for about the first hour and modest effort for some period beyond that. The first author is aware of legal cases in which plaintiffs were said to have made poor effort and yet, on a variety of neuropsychological measures, performed at levels comparable to preaccident testing.

Although posing practical difficulties, the recommendation has sometimes been made to intersperse measures of effort across evaluation sessions (e.g., Heilbronner et al., 2009), which could help to discourage overly global judgments. Moreover, in many circumstances, we lack the required scientific knowledge to determine the extent of generalization from low performance on an effort test to performance on various tests of ability. A poor result on a measure of effort may place results on other tests in question but of course will not establish unequivocally that they underrepresent ability or especially that functioning is intact. For example, a person who underperforms might also be injured or impaired. As we have emphasized throughout the chapter, these messier or more complex presentations remain less well understood and need to be investigated much more extensively.

Extreme Results of Malingering Tests Are Often Not What They Seem

Due to flaws in research designs or sampling methods, some malingering tests generate absurdly

extreme results (e.g., Mr. Smith's score falls 7.4 SD below the mean for an injured group). In a normal distribution, a *z* score of −5.0 occurs in less than 1 per 3,000,000 individuals, of −6.0 in about 1 per 800,000,000 individuals, and a score of −7.0 is infinitesimally small. Although various malingering tests do show strong features and are welcome additions to the field, these sorts of *z* scores should not be taken seriously because they are usually produced by skewed or distorted distributions and other methodological artifacts. As we have also described at length, due to the EGP, research often produces inflated accuracy rates or effect sizes. The concern is that these inflated results may be interpreted or presented literally, creating a gross misimpression about the strength of the evidence or the surety of malingering detection. Such practices treat plaintiffs unjustly, can lead to highly destructive consequences, and arguably should be flagged and strongly discouraged by the profession.

Response Set Measures on Questionnaires/Appraisal of Informants

Unlike measures of ability, on which individuals cannot intentionally perform better than they are able, faking good or dissimulation can occur on the evergrowing range of questionnaires used in clinical and forensic evaluations in neuropsychology. The development of questionnaires to measure such domains as everyday capacities and executive functions seeks to fill fundamental gaps within the field and can add important information to assessments. This is not an appropriate forum to address the strengths and weaknesses of such questionnaires comprehensively. We would simply note that subscales within questionnaires that are designed to detect over- or underreporting often have not been adequately studied. There seems to be a tendency at times to accept results on such response set measures almost at face value or by default, even if there is little or no research on the topic. Accurate measurement of response set can be demanding, and it may take extensive effort to evaluate, refine, and modify scales to reach modest or greater levels of validity. To assume such a positive accomplishment has been realized without scientific testing or on the basis of an isolated study or two can be wishful thinking. Additionally, for more traditional personality tests that are used in the context of neuropsychological evaluation, with the exception of the MMPI-2, the available literature on response sets is commonly inadequate or has generated mixed outcomes (see Rogers, 2008).

When conducting forensic assessments, it is often wise to seek information from collateral sources. Information gathering may involve interviews, the use of third-party (other) reporting forms that are available for various questionnaires, or both. It would seem prudent to use at least one method that provides a check on reporting tendencies, although again scientific foundations for assessing response set may be weak or practically nonexistent. Additional research and refinement of response set measures for both self-report and third-party versions of questionnaires and various personality tests used in neuropsychology would facilitate clinical and forensic evaluation. Furthermore, even if assessment methods are limited to interview, continuing efforts to develop structured procedures specifically aimed at neuropsychological issues (e.g., post-concussion symptoms) and which include appraisal of response sets could be very valuable.

Potential Benefits of Adaptive Testing

Concentrated efforts to explore the use of adaptive testing in malingering assessment (and neuropsychology in general) might prove fruitful. Adaptive testing offers the advantages of flexibility while potentially maintaining the types of formalization and scientific grounding that bolster decision accuracy. None of the authors doubt the potential value of flexibility or modification of procedures in relation to the questions at hand and initial testing results. Rather, the primary concern is with the methodology neuropsychologists use when implementing such an approach, which is often impressionistic or overly subjective and prone to many sources of judgment error.

For example, it would be interesting to sample self-reports of intact and impaired areas of functioning and, on that basis, determine the domains in which to perform forced-choice testing. Forced-choice techniques are highly malleable and can be designed for almost any content area, and thus fitting complaints with forced-choice procedures is feasible. A study might involve random assignment of content areas for forced-choice procedures, standardized content areas, and tailored content areas in relation to self-reports of functioning. Another approach would be to briefly sample impressions of test difficulty, look for discrepancies between impressions and true difficulty, and emphasize those areas in malingering assessment. A third approach would involve tailoring malingering assessment to areas of poor performance on standard neuropsychological tests, because it is here that the question of true vs. feigned deficit often becomes most relevant.

Creative Use of Simulation Designs

Simulation studies often become less valuable over time as research knowledge advances. However, simulation designs offer a number of advantages, in particular knowledge of true status and much greater control over level of effort and other variables. Given these advantages, variations on simulation designs might provide unique information, although in most cases such research falls mainly within the context of discovery and a good deal of further study is required to achieve sufficient verification. To provide a few examples of possible research directions, attempts could be made to appraise malingering skills and the ability to escape detection. One could then examine group differences to try to develop something akin to the MMPI-2 K Scale. For example, a group that can beat one or more malingering tests or fool clinicians may show other systematic differences (e.g., suppressed variation in test scores) compared to less successful malingerers or a group with true injury.

In simulation designs, one can also systematically manipulate degrees of effort across multiple levels, such as very high, to moderately high, to moderately low, to very low. Given greater variation in levels of effort, certain trends might appear that would otherwise be missed. Within-group designs might also prove informative. Although in many cases research on a new measure starts with simulation studies, it might be interesting to reverse the order at times and use simulation designs as consistency tests. For example, suppose based on retrospective case analysis or a contrasting group method that certain performance characteristics or signs are believed to indicate the joint presence of malingering and injury. One could follow up such work with simulation designs to see if the same findings hold. Of course, although consistency or inconsistency is far from definitive, in many areas of science a fundamental validation strategy is to appraise consistency in outcome across different methods for testing hypotheses or investigating phenomena.

Some Additional Thoughts on Research

Some approaches that seem promising for malingering detection represent an attempt to take commonsense considerations that many practitioners already apply impressionistically and place them on a more explicit, systematic, and formal basis to facilitate scientific testing and comparison. Ideally, the aim should be as much to verify adequate or good effort as it is to identify insufficient or poor effort.

Some disorders would seem to have relatively predictable outcomes. For example, with mild head injury, we would not expect catastrophic symptoms or a 6-month delay in symptom onset, and we would be much more likely to see problems in new learning rather than difficulties remembering major life events that occurred preinjury. If we could develop better measures of prototypical outcome and range of expected variation from prototypicality among those with genuine disorder, and if the level of variation was not too great, we would be in much better position to say that some outcome does not fall within expectations or is implausible. Such

measurements should be reducible to one or a few dimensions, with studies conducted to look at distributions among those with and without the disorder (including those feigning). One might call these types of measures indices of prototypicality. If outcome was so varied that most anything was about equally possible, it would serve as a more general warning about formulating causal judgments. A few words of caution are necessary here. We should be very careful about measures of severity, because one does not want to systematically identify those with genuine but atypically bad outcomes as malingerers. Also, failure to fit expectations for a particular type of injury only suggests that individuals do not have that type of injury, not necessarily that they are malingering – it may just be something else that ails them.

Some intentional symptom production requires constant attention. A patient who portrays a severe tremor may have difficulty doing so when fencing with the attorney on cross-examination. Using analogous approaches, we can examine what happens to intentionally produced symptoms under distracting conditions.

It may be possible to get at the intentionality of misrepresentations if we could create some index that compares the expected odds of misrepresentations working for or against the individual's self-interests and the examinee's obtained distribution. Some examinees misrepresent matters in a way that could cost them large settlement dollars. For example, some seriously impaired individuals deny problems, even when they have much to gain from accurate reporting. Other individuals show a very different pattern. For example, when it comes to remembering preinjury events, they seem to systematically forget most of their shortcomings but remember many of their strengths; the pattern is reversed when it comes to postinjury events, in which case they show remarkable recall of their shortcomings but seem to forget most of their accomplishments. Unintentional misrepresentations are not likely to work systematically in the direction of serving the person's legal case or self-interests. It would not seem that difficult a matter to derive methods for grading level of self-interest and classifying responses. Approaches that indicate deviation from expected patterns of error might be similarly useful in identifying when individuals have underrepresented their problems in a manner that could greatly impede fair resolution of their case.

A related index might measure negative consequences or events that have accrued for a person in proportion to the negative consequences claimed. Take an individual, for example, who reports intolerable pain but will not take a medication with mild side effects. One would expect some correlation between the level of suffering someone is experiencing and the level of suffering or inconvenience someone will tolerate in an effort to achieve improvement. The examinee who claims to be deeply distressed by being off the job but will not participate in a work-hardening program or even send out applications, has experienced no loss of income as a result of a generous benefit package, and has maintained an active recreational life, would seem much more likely to be a malingerer than the individual who has voluntarily undergone multiple painful operations, has had his house repossessed, and almost never goes out with friends. This type of index bears some resemblance to comparisons between subjective complaints and hard examination findings, although it is obviously problematic that some serious physical disorders or conditions often cannot be detected objectively. Therefore, it might be helpful to examine the relation between claimed distress and the level of negative consequences that have occurred or to which the individual has willingly submitted, such as reduction in income, pleasurable activities, and personal freedoms, and exposure to painful or dangerous medical procedures. Such indices might also consider what individuals have to gain if their legal cases are concluded in their favor.

In some situations it is to an individual's advantage to be (or appear to be) impaired, and in other (most) instances it is advantageous to be unimpaired. For example, if an individual is feigning paralysis of a limb to obtain a large settlement, a burning building can suddenly alter the contingencies. In the course of assessment, treatment, and day-to-day living, the relative balance

of incentive and disincentive for competence and impairment can shift dramatically, and in some circumstances individuals who have something to gain by being competent may not realize that their behavior could be detected or that they are falling out of role. Thus, the patient feigning neurological deficit suddenly becomes capable when appearing in a separate custody dispute, or an individual with severe spatial deficits instantaneously regains abilities when taking a driving examination. Other times matters are perhaps less obvious. The patient with supposed problems in word finding becomes articulate when needing to defend herself during cross-examination, or the individual who appears to struggle with the motoric aspects of writing signs the release form for the office secretary with good quality penmanship. It seems worthwhile to try to identify instances in which the contingencies for proficiency shift and to examine the extent to which levels of performance shift accordingly. Of course, as with other suggested indicators, the point is not merely to identify malingering but equally so to verify cooperation or lack of malingering.

The further development of procedures for assessing positive effort would be useful. One approach would be to obtain the best possible indicators of prior functioning, ideally in areas unlikely to be affected by the condition of interest and, even better, in areas that malingerers are likely to believe ought to be affected. One prefers measures of prior ability that were obtained in situations in which individuals would likely be motivated to do their best (e.g., preemployment ability testing). Based on these indicators, such as scores on past aptitude testing, one can predict level of performance. When these predictions are met or exceeded, one would have potentially strong evidence of adequate effort. As a simplified example, if someone who had obtained a Full Scale IQ score of 100 on a preinjury administration of a version of the Wechsler Intelligence Scale achieved a comparable score on postinjury testing, we would have good reason to assume that adequate effort was made on the test. Decreased scores are ambiguous, but the point of this procedure is not necessarily to identify inadequate effort, because we already have a variety of methods to do that, but rather to identify good effort.

Past indicators of ability, even those unlikely to be altered by the condition at issue, are fallible markers of postinjury abilities. The trick is to combine multiple fallible indicators properly (empirically and statistically) to construct stronger composites and to make predictions across a range of functions. One should be able to formulate error terms or distributions of expected results. We could then examine the match between expected and obtained results. For example, we might make predictions in five domains that should be unaltered by, say, mild to moderate head injury, and then look at the correspondence between the distribution of expected performance levels and that of obtained levels. In some cases at least, we might uncover powerful evidence of good effort. These methods might well turn out to have excellent valid-positive rates, giving us something roughly equivalent to symptom validity testing in the domain of good effort, that is, a procedure that more often than not yields evidence of limited use (related to low sensitivity), but one for which the value of the exceptions makes it well worthwhile.

We realize that a number of issues would need to be addressed (e.g., identifying the best predictors of later performance, difficulties interpreting performance that is lower than expected, identifying areas that are unlikely to be affected by injury), but we do not see these problems as insurmountable. The potential utility that measures of good effort would have for legal and nonlegal assessment would seem to warrant the attempt.

References

Ahern, D. C. (2010). *Extreme group comparisons: Nature, prevalence, and impact on psychological research*. Unpublished doctoral dissertation, University of Rhode Island, Kingston.

Baade, L. E., & Schoenberg, M. R. (2004). A proposed method to estimate premorbid intelligence utilizing group achievement measures from school

records. *Archives of Clinical Neuropsychology, 19*, 227–243.

Binder, L. M., Iverson, G. L., & Brooks, B. L. (2009). To err is human: "Abnormal" neuropsychological scores and variability are common in healthy adults. *Archives of Clinical Neuropsychology, 24*, 31–46.

Binder, L. M., & Rohling, M. L. (1996). Money matters: A meta-analytic review of the effects of financial incentives on recovery after closed-head injury. *American Journal of Psychiatry, 153*, 7–10.

Bridges, A. J., Faust, D., & Ahern, D. (2009). Methods for the evaluation of sexually abused children: Reframing the clinician's task and recognizing its disparity with research on indicators. In K. Kuehnle & M. Connell (Eds.), *The evaluation of child sexual abuse allegations* (pp. 21–47). Hoboken: Wiley.

Bridgman, P. W. (1927). *The logic of modern physics*. New York: Macmillian.

Brooks, B. L., Iverson, G. L., Sherman, E. M. S., & Holdnack, J. A. (2009). Healthy children and adolescents obtain some low scores across a battery of memory tests. *Journal of the International Neuropsychological Society, 15*, 613–617.

Brooks, B. L., Strauss, E., Sherman, E. M. S., Iverson, G. L., & Slick, D. J. (2009). Developments in neuropsychological assessment: Refining psychometric and clinical interpretive methods. *Canadian Psychology, 50*, 196–209.

Butcher, J. N., Graham, J. R., Ben-Porath, Y. S., Tellegen, A., Dahlstrom, W. G., & Kaemmer, B. (2001). *MMPI-2 (Minnesota Multiphasic Personality Inventory-2): Manual for administration, scoring, and interpretation, revised edition*. Minneapolis: University of Minnesota Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. San Diego: Academic.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.

Dawes, R. M., & Meehl, P. E. (1966). Mixed group validation: A method for determining the validity of diagnostic signs without using criterion groups. *Psychological Bulletin, 66*, 63–67.

Dumont, R., & Willis, J. O. (1995). Intrasubtest scatter on the WISC-III for various clinical samples vs. the standardization sample: An examination of WISC folklore. *Journal of Psychoeducational Assessment, 13*, 271–285.

Faust, D. (1997). Of science, meta-science, and clinical practice: The generalization of a generalization to a particular. *Journal of Personality Assessment, 68*, 331–354.

Faust, D. (2011). *Coping with psychiatric and psychological testimony* (6th ed.). New York: Oxford University Press.

Faust, D., & Ahern, D. C. (2011). Clinical judgment and prediction. In D. Faust, *Coping with psychiatric and psychological testimony* (6th ed.) (pp. 147–208). New York: Oxford University Press.

Faust, D., Ahern, D. C., & Bridges, A. J. (2011). Neuropsychological (brain damage) assessment. In D. Faust, *Coping with psychiatric and psychological testimony* (6th ed.) (pp. 363–469). New York: Oxford University Press.

Faust, D., Bridges, A. J., & Ahern, D. (2009a). Methods for the evaluation of sexually abused children: Issues and needed features for abuse indicators. In K. Kuehnle & M. Connell (Eds.), *The evaluation of child sexual abuse allegations* (pp. 3–19). Hoboken: Wiley.

Faust, D., Bridges, A. J., & Ahern, D. (2009b). Methods for the evaluation of sexually abused children: Suggestions for clinical work and research. In K. Kuehnle & M. Connell (Eds.), *The evaluation of child sexual abuse allegations* (pp. 49–66). Hoboken: Wiley.

Frederick, R. I., & Bowden, S. C. (2009). The test validation summary. *Assessment, 16*, 215–236.

Gough, H. G. (1954). Some common misconceptions about neuroticism. *Journal of Consulting Psychology, 18*, 287–292.

Greene, R. L. (2011). *The MMPI-2/MMPI-2-RF: An interpretive manual* (3rd ed.). Boston: Allyn & Bacon.

Heilbronner, R. L., Sweet, J. J., Morgan, J. E., Larrabee, G. J., & Millis, S. R. (2009). American Academy of Clinical Neuropsychology Consensus Conference statement on the neuropsychological assessment of effort, response bias, and malingering. *The Clinical Neuropsychologist, 23*, 1093–1129.

Hyman, R. (1977). "Cold reading": How to convince strangers that you know all about them. *The Zetetic, 1*, 18–37.

Kareken, D. A., & Williams, J. M. (1994). Human judgment and estimation of premorbid intellectual function. *Psychological Assessment, 6*, 83–91.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66*, 195–244.

Meehl, P. E. (1995). Bootstraps taxometrics: Solving the classification problem in psychopathology. *American Psychologist, 50*, 266–275.

Meehl, P. E. (1999). Clarifications about taxometric method. *Journal of Applied and Preventive Psychology, 8*, 165–174.

Meehl, P. E. (2001). Comorbidity and taxometrics. *Clinical Psychology: Science and Practice, 8*, 507–519.

Meehl, P. E. (2004). What's in a taxon? *Journal of Abnormal Psychology, 113*, 39–43.

Orme, D., Ree, M. J., & Rioux, P. (2001). Premorbid IQ estimates from a multiple aptitude test battery: Regression vs. equating. *Archives of Clinical Neuropsychology, 16*, 679–688.

Reynolds, C. R. (1997). Postscripts on premorbid ability estimation: Conceptual addenda and a few words on alternative and conditional approaches. *Archives of Clinical Neuropsychology, 12*, 769–778.

Reynolds, C. R. (1998). Common sense, clinicians, and actuarialism in the detection of malingering during head injury litigation. In C. R. Reynolds (Ed.), *Detection of malingering during head injury litigation* (pp. 261–286). New York: Plenum.

Rogers, R. (1990a). Development of a new classificatory model of malingering. *Bulletin of the American Academy of Psychiatry and Law, 18*, 323–333.

Rogers, R. (1990b). Models of feigned mental illness. *Professional Psychology: Research and Practice, 21*, 182–188.

Rogers, R. (Ed.). (2008). *Clinical assessment of malingering and deception* (3rd ed.). New York: Guilford.

Rogers, R., Sewell, K. W., Martin, M. A., & Vitacco, M. J. (2003). Detection of feigned mental disorders: A meta-analysis of the MMPI-2 and malingering. *Assessment, 10*, 160–177.

Schretlen, D. J., Buffington, A. L. H., Meyer, S. M., & Pearlson, G. D. (2005). The use of word-reading to estimate "premorbid" ability in cognitive domains other than intelligence. *Journal of the International Neuropsychological Society, 11*, 784–787.

Schretlen, D. J., Munro, C. A., Anthony, J. C., & Pearlson, G. D. (2003). Examining the range of normal intraindividual variability in neuropsychological test performance. *Journal of the International Neuropsychological Society, 9*, 864–870.

Vickery, C. D., Berry, D. T. R., Inman, T. H., Harris, M. J., & Orey, S. A. (2001). Detection of inadequate effort on neuropsychological testing: A meta-analytic review of selected procedures. *Archives of Clinical Neuropsychology, 16*, 45–73.

Waller, N. G., & Meehl, P. E. (1998). *Multivariate taxometric procedures: Distinguishing types from continua*. Thousand Oaks: Sage.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale, fourth edition: Administration and scoring manual*. San Antonio: The Psychological Corporation.

Williams, J. M. (1997). The prediction of premorbid memory ability. *Archives of Clinical Neuropsychology, 12*, 745–756.

Williams, J. M. (1998). The malingering of memory disorder. In C. R. Reynolds (Ed.), *Detection of malingering during head injury litigation* (pp. 105–132). New York: Plenum.

Wong, J. L., Regennitter, R. P., & Barris, F. (1994). Base rates and simulated symptoms of mild head injury among normals. *Archives of Clinical Neuropsychology, 9*, 411–425.

http://www.springer.com/978-1-4614-0441-5

Detection of Malingering during Head Injury Litigation
Reynolds, C.R.; Horton, Jr., A.M. (Eds.)
2012, XII, 381 p., Hardcover
ISBN: 978-1-4614-0441-5