Nova Norma de setembro-outubro de 2004

Relatório preliminar de mudanças no processo de normatização e de correção da nova norma do Sigma Test em relação à norma de setembro de 2003.

1 – Em 2003 a dificuldade era calculada para cada questão, usando a fórmula ***D=(E/(1-E))***, onde ***D*** é a dificuldade, ***E*** é o número de pessoas que erraram a questão. Agora a dificuldade é calculada para cada resposta. Isso torna os escores mais acurados. Por exemplo: na questão 35 há apenas 1 pessoa que deu resposta totalmente certa e 6 pessoas que deram respostas incompletas. Se mais uma pessoa fizer o teste e der resposta incompleta, então, pelo método de 2003, a dificuldade da questão inteira diminuiria, de modo que tanto as pessoas que deram resposta incompleta quanto a pessoa que deu a melhor resposta receberiam menos pontos. Pelo novo método, só as pessoas que deram resposta incompleta receberão menos pontos. Um caso ilustrativo: havia 144 *testees*, sendo que um deles deu resposta certa na questão 35, enquanto 6 deram resposta incompleta nesse item (e os outros 137 deram respostas erradas). Mais uma pessoa fez o teste e deu resposta parcialmente certa na questão 35. Pelo método de 2004, o peso da resposta certa era 143 e subiria para 144, enquanto o peso da resposta incompleta cairia de 137/7 para 137/8. A proporção poderia ser calculada de várias maneiras mais sofisticadas, como a proporção entre a média aritmética dos escores brutos das pessoas que deram a resposta certa e das pessoas que deram cada resposta incompleta. Mas nos casos em que poucas pessoas deram determinada resposta, esse sistema ficaria comprometido.

2 – As probabilidades de a pessoa acertar cada questão eram calculadas usando um parâmetro constante para todas as questões. Essa era uma hipótese ruim, porque a curva que descreve a probabilidade de acerto de cada questão em função do nível de habilidade é diferente de uma questão para outra. Isso não interfere no escore final, mas possibilita determinar mais um escore para o teste: "***QI baseado em escore fracionado***", que é o escore bruto mais provável de ser alcançado por uma pessoa que tenha determinado escore ponderado. Em seguida, esse escore bruto é estandardizado e convertido em escore com média 100 e desvio-padrão 16. Esse é o escore que mais se assemelha ao método da norma antiga (antes de setembro de 2003), com coeficiente de correlação linear de Pearson 0,966 (com base em 145 pares de escores) entre esses escores fracionados e os escores da norma antiga (anterior a setembro de 2003).

Ainda serão implementadas mudanças no método atual, tais com:

Escore ***g***", com base na pontuação bruta multiplicada pela carga fatorial ***g*** de cada questão. É importante ter em mente que essa carga fatorial ***g*** é uma hipótese "capenga", porque o fator comum medido pelas questões do teste não é necessariamente o fator ***g*** de Spearman. No caso de um teste cultural, por exemplo, haverá um fator comum para todas as questões, no entanto esse fator nada tem a ver com o ***g*** de Spearman. Analogamente, o fator comum de um teste cognitivo também estará impregnado com uma certa dose de cultura e outros "ruídos", não será um escore ***g*** puro, embora seja comparativamente mais puro do que o escore bruto que atribui carga 1 a cada questão. Enfim, será uma informação complementar ao teste, mas que deve ser interpretada com reservas.

O nível de raridade é uma informação tradicionalmente presente em relatórios de contagem, mas é de pouco valor prático pelo fato de a função que descreve a distribuição real de freqüência degringolar fora do intervalo de -2σ a $+2\sigma$. Em Giga Society, por exemplo, o corte teórico é de 1 em 1.000.000.000, mas o corte verdadeiro não é maior do que 1 em 100.000. O mesmo problema afeta Sigma VI e outras sociedades que pretendem discriminar em níveis acima de 4 desvios-padrão usando escalas baseadas em raridade. No caso da Mensa, o corte teórico é bem semelhante ao verdadeiro, porém os traços latentes medidos pelo RAPM e pelo Cattell são

excessivamente dependentes da velocidade para solucionar problemas simples, o que não reflete a capacidade para lidar com problemas profundos e complexos e não é uma representação satisfatória da inteligência para altos níveis de desempenho. É aceitável para percentil 98, mas não para percentis muito acima desse ponto.

No processo de correção do Sigma Test, a única mudança é a seguinte: em vez de colocar escores 1 ou fração para cada resposta, passamos a colocar 1 ou letras de A até E. Quando for zero, então usamos letra Z. Por exemplo:

Questão n solução ótima: 1
Questão n solução muito boa: A
Questão n solução boa: B
Questão n solução precária: C

Um exemplo ilustrativo: calcular a velocidade da luz pelo método de Foucault recebe ponto inteiro (1), pelo método de Fizeau recebe ponto A, pelo método de Roemer recebe ponto B, pelo método do forno de microondas também recebe ponto B. Se 3 pessoas resolveram pelo método de Foucault, cada uma receberá 142/3 pontos. Se 10 pessoas resolveram pelo método de Fizeau, cada uma receberá 132/13 pontos etc.

As questões com correções problemáticas 29, 30 e 36 passaram a receber escores fáceis de calcular e mais justos, com base na segunda melhor solução. Se em algum momento alguém encontrar a melhor solução, esta imediatamente será incorporada e o teste continuará calibrado. Pelo método de setembro de 2003, quando ninguém dava resposta totalmente certa a uma questão, ela ficava com peso indefinido (divisão por zero). Agora ela fica temporariamente com o peso determinado pela segunda melhor resposta, ou pela terceira melhor resposta etc. No caso da questão 29, se apenas uma pessoa deu resposta muito boa, que não é totalmente certa, mas é melhor do que as outras, essa pessoa receberá 144 pontos, porque o nível de dificuldade dessa segunda melhor resposta é equivalente (pela raridade de acertadores) ao de uma solução 100% que tenha 1 acertador. Isso mantém a ponderação mais coerente do que o método anterior.

Pelo método de setembro de 2003, havia uma planilha específica para cada questão problemática. Agora elas podem entrar diretamente com as outras questões no cálculo dos pesos e gerar valores mais coerentes com os das demais questões, além de tornar o processo de correção mais ágil e mais fácil.

Os diferentes escores na questão 36, que estavam todos 0, agora podem receber pontos diferenciados e contribuir para melhorar a precisão do escore final.

Há uma maneira interessante de calcular a incerteza nos escores: a diferença entre os pontos brutos obtidos e os pontos brutos esperados com base no escore ponderado. Para cada questão há uma probabilidade P de a pessoa acertar. A soma dos quadrados das diferenças entre a fração do ponto obtido e a probabilidade de marcar aquele ponto cheio, dividido pelo número de questões menos 1 deve dar satisfatoriamente uma variância, e a raiz dessa variância é um valor razoável para representar a incerteza no escore final. Outro método para calcular a incerteza é o índice de informação da TRI, que provê incertezas para escores brutos, mas pode ser ajustada para escores ponderados. A vantagem em relação ao método da TRI tradicional é que podemos determinar a incerteza no escore de cada pessoa, em vez de determinar a incerteza em cada escore (independente de quais questões geraram aquele escore), o que faz uma grande diferença e torna esse método muito mais acurado do que o usado em TRI. Em todos os casos, estamos falando de **incerteza no escore**, que não é o mesmo que **incerteza no QI "verdadeiro" correspondente a esse escore**. Ou seja: **rIQ** = 150,0 $\eta_{(+0,9)(-1,3)}$ significa que há 50% de chances de o escore no teste estar entre 148,7 e 150,9. Isso não significa que o QI verdadeiro tem essa mesma incerteza. Um

procedimento muito comum é calcular a incerteza levando em conta a raridade, assim, para QIs acima de 100, a incerteza para cima sempre será menor do que para baixo. Isso é válido para o real potencial intelectual, mas não para o número que representa esse potencial. O método usado no Sigma Test consiste em calcular a incerteza com base na raiz quadrada da soma dos quadrados das diferenças entre a pontuação obtida em cada item e a pontuação esperada no respectivo item. Quando a pontuação esperada é maior do que a obtida, isso indica erro para menos (o escore nesse item foi menor do que o "correto"). Quando a pontuação obtida é maior do que a esperada, isso indica erro para mais (o escore nesse item foi maior do que o "correto").

Legenda sobre os escores do certificado

Potential IQ: *indica proporção de potencial. Um QI 132 indica nível de produção intelectual equivalente a 19 pessoas com QI 100 somadas (porém não em sinergia), enquanto um QI 148 indica nível de produção intelectual equivalente a 84 pessoas com QI 100 somadas (porém não em sinergia).*

Potential IQ gauged: *é calibrado com base na diferença de incerteza para cima e para baixo. Se a incerteza para cima for maior, o QI calibrado será maior, caso contrário será menor. A calibragem consiste simplesmente em somar 1,48η (=1σ) para cima e subtrair 1,48η (=1σ) para baixo. Basicamente essa calibração atenua o efeito produzido nos casos de pessoas que conquistam quase todo o escore em poucas questões, contrabalançando o fato com um aumento proporcional nos escores das pessoas acertaram maior quantidade de questões, de modo que as médias aritmética de todos escores antes e depois da calibração fiquem iguais.*

Rarity IQ: *indica nível de raridade, semelhante ao sistema Wechsler de padronização, porém com desvio-padrão 16. Um QI 132 indica percentil 97,7, enquanto um QI 148 indica percentil 99,87. É uma escala de intervalo de QI que representa bem uma escala de proporção de raridade. Essa raridade tanto pode ser em escores ponderados (rarity IQ based on balanced score), como em escores brutos (rarity IQ based on raw score) ou em escores fracionados (rarity IQ based on fractional score).*

Balanced Score: *indica a pontuação total alcançada no teste, atribuindo pontos diferenciados a cada questão, em função da raridade de acertadores. Uma questão com 10% de acertadores tem dificuldade 9, sendo D=(1-A)/A, em que "D" representa a dificuldade e "A" corresponde ao número de acertadores. Uma questão com 80% de acertadores tem dificuldade 0,25 e uma questão com 50% de acertadores tem dificuldade 1. Quando ninguém encontra a resposta certa, então o critério do peso recai sobre a segunda melhor resposta: Se ninguém acerta e 10% encontraram a segunda melhor resposta, então a segunda melhor resposta tem dificuldade 9. Quando algumas pessoas acertaram a melhor resposta e outras dão respostas com qualidade B, C, D, respectivamente:*

D claramente melhor do que resposta errada = 18% de pessoas
C claramente melhor do que D = 12% de pessoas
B claramente melhor do que C = 15% de pessoas
Resposta certa claramente melhor do que B = 5% de pessoas

Então a resposta certa recebe 19 pontos = 95/5.
A resposta com qualidade B recebe 4 pontos = (1-0,15-0,05)/(0,15+0,05).
A resposta com qualidade C recebe 2,125 pontos = (1-0,15-0,05-0,12)/(0,15+0,05+0,12).
A resposta com qualidade D recebe...
1 ponto = (1-0,15-0,05-0,12-0,18)/(0,15+0,05+0,12+0,18).

Raw Score: *indica a pontuação bruta alcançada no teste, atribuindo 1 ponto para cada resposta totalmente certa e 0 para qualquer resposta diferente da certa. O QI baseado no Raw Score é obtido calculando a média e o desvio-padrão dos Raw Scores e convertendo-os em QIs com mesma média e desvio-padrão do Balanced Score do grupo de pessoas examinadas. Se o grupo de pessoas examinadas teve QI médio 149,9 e desvio-padrão 14,6, e esse mesmo grupo teve Raw Score médio 20,38 com desvio-padrão 4,54, então uma pessoa com Raw Score 30 terá seu QI baseado no Raw Score igual a 180,8.*

Fractional Score: *indica a pontuação bruta esperada para uma pessoa que tenha alcançado determinado Balanced Score. Se uma pessoa tem QI 130 baseado no Balanced Score, ela tem 50% de chances de acertar questões com nível de dificuldade 130 e é esperado que acerte metade das questões desse nível. Quando o nível de dificuldade da questão é diferente do nível de habilidade da pessoa, quanto maior for o nível de habilidade da pessoa, tanto maiores são as chances de ela acertar aquela questão, mas a variação não é igual para todas as questões. Depende de um parâmetro de discriminação (a) e um parâmetro de dificuldade (b). O método será explicado com detalhes em breve. Por enquanto, quem tiver interesse em conhecer mais sobre isso, pode ler este artigo: http://www.sigmasociety.com/artigos/fuvest_2004_artigo.pdf e visitar alguns links citados nesse artigo, especialmente os que tratam de TRI.*

Percentil = Gaussiana cumulativa de (QI-100)/16

Quotiente of Potential (QP) = Proporção de Potencial entre A e B = e^[(QI$_A$-QI$_B$)/k], onde k≅10,828

pIQ = rIQ + 2,023 × e^[(rIQ-100)/32] ***pIQ*** *= Potential IQ* ***rIQ*** *= Rarity IQ*

*Para converter escore bruto em QI de raridade (**rIQ**) e ter uma idéia aproximada do QI correspondente a cada número de pontos brutos, use a seguinte fórmula:*

rIQ ≅ S x 2,84 + 92

*Onde **S** é o escore bruto (1 ponto para cada resposta certa). Se você está seguro de ter acertado 25 questões, então seu provável rIQ é 25 x 2,84 + 92 = 163. Isso corresponde a um pIQ 177. Se seu escore bruto fosse 21, seu rIQ seria 152. Com esse rIQ você poderia se associar em ISI-S ou Glia. Se seu escore bruto fosse 19, seu rIQ seria 146 e seu pIQ seria 155. Com esse rIQ você **não** poderia se associar em ISI-S ou Glia, mas poderia se associar em Sigma III, porque Sigma III utiliza o pIQ.*

*A maioria das sociedades de alto QI usa o rIQ **teórico** nos critérios para admissão, enquanto Sigma utiliza o pIQ. Na prática, muitas sociedades utilizam implicitamente escores muito semelhantes ao pIQ, como nos casos de Giga e Mega, isso porque os escores no Mega Test e no Titan Test são muito mais semelhantes a pIQ do que a rIQ. No entanto, embora usem escores semelhantes a pIQ, afirmam usar rIQ e com isso são cometidos alguns erros graves, como atribuir a Giga Society o nível de raridade de 1 em 1.000.000.000, quando o correto seria cerca de 1 em 100.000. No caso de Mega acontece o*

mesmo, porém com menor distorção (1 em 100.000 em vez de 1 em 1.000.000). Esses fatos não são novidades e há bastante tempo Kevin Langdon e Bob Seitz encontraram testos em torno de 1 em 100.000 para o Mega Test, com base em diferentes métodos de normatização. Isso situaria o cut-off verdadeiro de Mega Society em algo perto de 1 em 30.000.

A vantagem de usar escores pIQ é principalmente conceitual. O rIQ só faz sentido até percentil 98 ou no máximo 99, enquanto pIQ é bem definido para QI até 190 ou 200. Ou seja, o rIQ 137 provavelmente indica um verdadeiro nível de raridade de 1 em 100, mas um rIQ 150 pode indicar um verdadeiro nível de raridade muito diferente de 1 em 1.000, e essa distorção cresce para QIs mais elevados. Um escore rIQ 176 no Mega Test muito provavelmente não significa raridade 1 em 1.000.000, no entanto um escore pIQ 176 no Sigma Test significa corretamente a proporção de potencial definida para esse QI. A vantagem, portanto, está na possibilidade de interpretação do escore, atribuindo significados legítimos a cada pontuação, ao contrário do que acontece no caso do rIQ, que atribui raridades nitidamente inverossímeis aos escores, devido à forte deterioração da gaussiana fora do intervalo entre -2σ e +2σ.

Para conhecer mais detalhes sobre o Sigma Test e o processo de padronização, veja esta página: http://www.sigmasociety.com/sigma_teste.asp