Sigma Test – Norma de setembro de 2003

Revisado em outubro de 2004
Número de testees: 123 (22/08/2003)

É com grande satisfação que apresentamos o relatório de setembro de 2003 sobre o Sigma Test. Aproveito para agradecer a todos que traduziram o teste, que ajudaram na divulgação e que participaram diretamente na amostragem (submetendo planilhas de respostas). Graças a vocês, agora podemos experimentar esse método. Espero que a descrição do processo usado para normatização ajude a outros designers de testes, que também estejam enfrentando dificuldades com pequeno número de testees. O método permite calcular normas mais acuradas, mais pertinentes e com base em pequena quantidade de testees.

Parte das informações sobre história dos testes de inteligência e sobre conceitos de QI, que constava no texto original sobre a nova norma, foi revisada e desmembrada em artigos, que podem ser baixados aqui:

http://www.sigmasociety.com/artigos/historia.pdf (história dos testes)
http://www.sigmasociety.com/artigos/norma_set_2004.pdf (norma atualizada em outubro de 2004)
http://www.sigmasociety.com/artigos/fuvest_2004_artigo.pdf (artigo envolvendo nossa metodologia em conjunto com TRI)
http://www.sigmasociety.com/artigos/certificate_model.pdf (modelo de certificado)

➢ Conceitos de QI de potencial (pIQ):

Como todos sabemos, não é possível determinar o QI de maneira tão direta como se determina a altura ou a massa. Por isso o método que proponho requer alguma engenhosidade, para que possamos interpretar satisfatoriamente os elementos envolvidos e tratá-los sem que sejam introduzidas novas distorções, sob o pretexto de eliminar distorções antigas.

Começaremos pelo conceito de proporção: se uma pessoa tem QI 100 e outra tem QI 130, é evidente que não se pode dizer que a pessoa com QI 130 é 1,3 vezes mais inteligente ou 30% mais inteligente. Se uma pessoa com pIQ 100 acerta 7 questões entre um total de 30 num teste de inteligência, e outra pessoa com pQI 130 marca 18 num total de 30 no mesmo teste, isso não diz nada sobre o potencial comparativo dessas duas pessoas. Com base nesses dados, seria muito errado dizer que uma pessoa com pIQ 130 é 18/7 vezes mais inteligente do que uma pessoa com pIQ 100. Contudo, se 10 pessoas com pIQ 100, trabalhando independentemente no mesmo conjunto de problemas, conseguem resolver 18/30, e uma pessoa sozinha, com pIQ 130, também consegue marcar 18/30, podemos dizer que o potencial de uma pessoa com pIQ 130 é 10 vezes maior do que o de uma pessoa com pIQ 100, ou que uma pessoa com pIQ 130 produz intelectualmente tanto quanto 10 pessoas com pIQ 100 somadas. Se uma pessoa com pIQ 160 resolver 23/30, podemos esperar que 10 pessoas com pIQ 130 somadas também possam resolver 23/30; ou 100 pessoas com pIQ 100 também resolveriam 23/30. Apesar do aspecto simplista, essa hipótese é bem fundamentada e pode ser extensivamente confirmada com base em mais de 70.000 jogadores de Xadrez, *rankeados* pela FIDE e produzindo resultados estatísticos (2.300.000 jogos) desde 1971. Sempre ocorrem pequenas distorções, mas o método tem se mostrado excelente e claramente superior a qualquer sistema de normatização usado por psicometristas. Contudo, conforme veremos mais adiante, os escores baseados em rarity-IQ falham nestas predições, e se 10 pessoas com rIQ 100 produzem tanto quando 1 pessoa com rarity-IQ 130, a mesma proporção não se mantém quando passamos de 130 para 160. Em vez disso, uma pessoa com rIQ 160 produz mais do

que 10 pessoas com rIQ 130. O motivo disso, conforme veremos antes do final desse artigo, é que o pIQ não é bem representado pelo rarity-IQ exceto num estreito intervalo.

O termo pIQ que estamos introduzindo é uma grandeza logarítmica que representa a capacidade intelectual. A diferença entre os pIQs indica a proporção entre os níveis intelectuais: $P_1/P_2=e^{(k*\Delta QI)}$. Na época em que desenvolvi o método, meados de 2000, tomei como referências os artigos de Bill McGaugh e Grady Towers, bem como os trabalhos de Arpad Elo e John Scoville, mas o próprio Elo se baseou em Rasch, a quem deve ser atribuída a equação $P_1/P_2=e^{(k*\Delta QI)}$, fato que tomei conhecimento em 2004, quando comecei a trabalhar na Casa do Psicólogo e aprender sobre Teoria de Resposta ao Item (TRI), que se vale de muitos métodos semelhantes aos que desenvolvi, alguns melhores do que os meus e outros que podem ser aprimorados com base em minhas inovações. O assunto TRI é pouco difundido e é provável que Towers tenha escrito seu artigo sem conhecer os trabalhos de Rasch (e sucessores) que determinam o valor de k, porque em seu texto ele afirma que o valor de k não era conhecido, quando na verdade já se sabia o valor de k há várias décadas. O próprio Elo, duas décadas antes de Towers, se baseou em Rasch para formular o sistema de rating de Xadrez. De qualquer modo, mesmo que o valor de k não fosse conhecido, seria possível calcular essa grandeza sem necessidade de levantamentos estatísticos sobre testes, com base nos artigos de Towers e McGaugh. Foi o que fiz em 2000 e, mais tarde, em setembro de 2003, repeti com base em dados empíricos do Sigma Test, com resultados muito semelhantes.

Portanto a contribuição que demos com a metodologia apresentada aqui é menor do que havia sido dito em 2003, já que o valor de k era conhecido desde a década de 1930. Minhas contribuições em 2003 foram apresentar um conceito adequado para proporção de potencial e uma maneira mais adequada para calcular dificuldades. Em 2004, minhas contribuições foram determinar as variações do parâmetro c (de acerto casual) em função da habilidade, calcular os parâmetros a, b e c sem ajuste simultâneo, determinação das incertezas nos escores individuais e outros detalhes que serão descritos em breve, num trabalho exclusivamente sobre esse tema.

Uma proposta anterior para representar a relação P_1/P_2 foi apresentada por McGaugh, como representando a diferença de velocidades. Essa é a primeira idéia que vem a mente, mas não é adequada porque a correlação entre inteligência e velocidade é baixa e com grandes variações. Mas é uma informação interessante para calibrar escores de pessoas que resolvem testes em tempo menor do que o prazo. Para cada dobra, há um incremento em torno de 7 pontos, assim, se uma pessoa resolve em 12 minutos um teste com prazo de 40 minutos, e obtém escore 140, esse escore deve ser acrescido com 12 pontos e o QI dessa pessoa deve ser 152, porque $[ln(40/12)/ln(2)]*7\cong12$. Muito pior seria dizer que uma pessoa com QI 160 tem potencial duas vezes maior do que uma pessoa com QI 80. A maneira como interpreto a relação P_1/P_2 é provavelmente uma das mais adequadas: ***"Quantas pessoas com $pIQ_2=x$ são necessárias para atingir o mesmo nível de produção intelectual de outra pessoa com $pIQ_1=y$ $(y>x)$"?*** Em especial, interpreto a "produção intelectual" (nesse caso) como a capacidade de obter um determinado número de respostas certas num bom teste projetado para avaliar o nível de desempenho mental, mas o conceito pode ser estendido a outras atividades que envolvem produção intelectual.

O que eu chamo "bom teste de desempenho mental" precisa reunir determinadas características que estão presentes em alguns dos testes mais famosos que não usam limite de tempo. Essas características são: questões com vários níveis de dificuldade, que exigem diferentes tipos de pensamento, cujo teto de dificuldade seja compatível ao teto de QI que se deseja medir, que não sejam sobrecarregados culturalmente etc. Um teste com dezenas de questões muito fáceis, como o WAIS, Stanford-Binet, Cattell ou RAPM, por exemplo, pode ser apropriado apenas para pIQs até 135-140. Acima deste nível, os testes medem apenas quais

das pessoas com pIQ acima de 140 são mais velozes para solucionar problemas simples, mas não servem, por exemplo, para diferenciar entre pessoas com pIQ 140 e 150. Os testes podem dizer que as pessoas que marcam 25/30 surgem com freqüência 1 em 1000, e as pessoas que marcam 20/30 surgem com freqüência 1 em 200, mas isso não significa que quem marca 20/30 tem QI 140 e quem marca 25/30 tem QI 150, porque acima do nível 140 o teste está avaliando quais pessoas com 140+ são mais rápidas, e, embora a rapidez correlate positivamente com a inteligência, não é uma correlação forte. Então a pessoa que marcou 25/30 pode realmente ser alguém com QI 150, mas também pode ser alguém com QI 140 cuja rapidez é maior que 80% das outras pessoas com QI 140, e nesse caso ela ficará nos 20% superiores (1 em 1000), enquanto aquela que marcou 20/30 pode ter QI 140 ou pode ter 150, 160, 170 etc. e sua velocidade estar na média o grupo de 140 ou um pouco abaixo. Essas distorções não acontecem apenas no nível 140+, mas em praticamente qualquer nível de QI. O motivo pelo qual estamos comentando especificamente o problema no nível 140+, é que a partir daí as distorções se tornam muito maiores, porque chega ao ponto em que o pensamento profundo começa a ser um elemento mais importante na determinação da inteligência do que a velocidade de raciocínio.

Por isso, um teste cuidadosamente normatizado, com base em milhões de testees, capaz de discriminar num nível de 1 em 40.000.000 (rIQ 187), como é o caso do Cattell III, não tem nenhuma validade fora do intervalo de QI 60-140, a menos que as questões sejam suficientemente difíceis. E sabemos que as questões não são difíceis o bastante, portanto nenhum escore acima de 140 tem qualquer validade se tiver sido obtido em testes cronometrados. Conseqüentemente, tais testes não servem aos propósitos descritos aqui e não permitem estabelecer proporções de potencial, a não ser no intervalo entre 60 e 140.

➢ Sobre o processo de correção:

O método descrito a seguir foi proposto em setembro de 2003. É inadequado, mas é importante conhecê-lo para compreender melhor o método usado a partir de outubro de 2004.

Pela nova norma (set. 2003), as questões 1 até 25 podem receber “certo” ou “errado”. Se houver um descuido evidente, a pessoa não perde nenhum ponto. Vejamos alguns exemplos de descuidos evidentes: se a pessoa anotar a resposta para a questão 15 no lugar da questão 14 e vice-versa, isto é considerado descuido. Se a pessoa indicar 7/11 e em seguida representar da forma decimal como 0,77777..., isto é considerado descuido. Se a pessoa indicar 1000 onde deveria indicar 100, isso não é descuido, isso é considerado ERRO. Só será considerado descuido quando houver alguma evidência consistente de que a pessoa resolveu corretamente, chegou na solução final certa, porém se equivocou ao anotar a resposta, ou algo assim. Ou se a pessoa anotou incorretamente um dado fornecido pelo problema e fez toda a solução usando o número errado, mas com o procedimento certo. Isto também é considerado descuido.

Para as questões 26 até 36, nos casos em que existe mais de uma solução aceitável, se a resposta estiver 100% certa, receberá o ponto ponderado inteiro. Se não for 100% certa, então receberá uma fração de ponto proporcional à freqüência com que a resposta acontece. Por exemplo: a questão 35 tem uma solução ótima que só foi encontrada por uma pessoa e tem diversas soluções aceitáveis, porém não otimizadas, que foram encontradas por 5 pessoas. Além de zilhões de soluções erradas. Então o peso da resposta ótima é 6 vezes maior do que o peso das respostas não-otimizadas. Em outras palavras, as respostas não otimizadas recebem 1/(5+1) do ponto conferido à resposta plena. No caso da questão 31, tivemos 12 respostas certas e 12 respostas erradas com erro pequeno. Então a resposta certa recebe o ponto cheio e essas respostas com erro pequeno recebem 0,5 (12/24). No caso da questão 33,

existe uma solução certa, que 2 pessoas encontraram, uma solução parcialmente certa, que 4 pessoas encontraram, e uma solução errada (que poderia funcionar, não fosse por um detalhe), que 42 pessoas encontraram e também recebe uma parte do ponto para diferenciá-la das soluções completamente erradas. Então as 2 pessoas que deram a resposta certa receberam o ponto cheio. As 4 pessoas que deram a resposta parcialmente certa recebem 2/6 (1/3) do ponto e as 42 pessoas que deram a resposta errada (com erro pequeno) recebem 2/48 (1/24) do ponto (48=2+4+42). Quando uma questão admitir mais do que um método igualmente eficaz, como na questão 27, que admite várias soluções aproximadamente igualmente eficientes, então cada uma delas receberá 100% dos pontos (desde que seja completa). As frações de pontos são aplicadas exclusivamente nos casos em que existem diferenças de qualidade ou abrangência entre as diferentes soluções. Em tais casos, as soluções incompletas recebem uma fração de ponto proporcional à relação entre a quantidade de respostas completas e incompletas. O mesmo sistema se aplica em todas as questões entre 26 e 36. Para os testes que admitem mais do que uma solução e existem diferenças qualitativas entre as respostas, o sistema de frações contribui para aumentar a acurácia do teste. (veja esse método aprimorado nesse artigo: http://www.sigmasociety.com/artigos/norma_set_2004.pdf).

Essa distinção entre as questões 1-25 e as outras é fundamental para que o teste atenda ao objetivo de aferir corretamente a capacidade intelectual. Cada designer terá que fazer um exame crítico e identificar as particularidades de seu teste, reconhecendo as questões que exigem atenção especial e tratando-as adequadamente, aplicando ajustes similares aos apresentados acima. A falta de versatilidade é um dos grandes problemas dos psicometristas que insistem em usar métodos clássicos, sem atentar para as sutilezas que podem exigir tratamento personalizado. Isso não é adulteração. Isso é calibração, é refinamento. E mais adiante veremos como fazer a sintonia fina dessa calibragem. Qualquer uso da força bruta, tentando usar o mesmo critério para todas as questões, seria acrítico e produziria resultados menos acurados.

➢ Sobre os pesos das questões:

O peso de cada questão é determinado da seguinte forma: ***P=Ne/Nc***, onde ***P*** é o peso, ***Ne*** é o número de respostas erradas e ***Nc*** é o número de respostas certas (inclusive as frações). No caso da questão 35, por exemplo, que teve apenas um acertador e 5 soluções não otimizadas (valendo 1/6 cada), o peso é ~66, porque dos 123 testees tivemos 1 resposta certa e 5 respostas valendo 1/6 do ponto, ou seja: 1,8333... de certos e 121,166... de erros. Então 121/1,83 ~66. A pessoa que acertou a questão, recebeu 66,09 pontos e as 5 pessoas que deram respostas parcialmente certas, receberam 11,01 cada (veja esse método aprimorado nesse artigo: http://www.sigmasociety.com/artigos/norma_set_2004.pdf).

A questão 1, que todos acertaram, tem peso 0, ou seja, ela é indiferente para todos os testees. A partir do momento que alguém errar esta questão, ela passará a somar pontos para todos aqueles que a tiverem acertado. Digamos que o próximo testee erre esta questão. Então ela terá peso 1/(124-1), ou seja: 0,008. A questão 2, que teve 3 erros e 120 acertos, tem peso ~0,025. A questão 36 tem peso maior do que 245 (porque entre 123 pessoas, ninguém enviou uma solução 0,5 certa).

Com este método simples, podemos praticamente anular o problema dos descuidos. Algumas vezes, uma pessoa (A) responde a um teste com 30 questões, e por dedicar pouca atenção às questões mais fáceis, acaba cometendo erros que não são representativos de sua real capacidade. Ela demonstra isso resolvendo algumas das questões mais difíceis. No entanto, se ela cometer 5 descuidos nas questões fáceis e acertar todas as outras, ela terá escore 25, e se uma outra pessoa (B) prestar mais atenção ao teste e resolver as questões mais fáceis com o

maior cuidado, e deixar de resolver as duas mais difíceis, porque não consegue, então a pessoa B vai marcar 28 pontos. Qual destas pessoas é a mais inteligente? Qual é a mais descuidada? Eu diria que a pessoa "A" é mais inteligente e também mais descuidada. No entanto, os testes que não usam sistemas ponderados na avaliação, diriam que a pessoa "B" é mais inteligente. Com os pesos calculados conforme descrito acima, esse problema desaparece, e a pessoa objetivamente recebe uma avaliação representativa de sua capacidade, mesmo que ela cometa 5 ou 10 descuidos. Por exemplo: uma pessoa (A) que acertar a questão 35 do Sigma Test, que é muito difícil, e 20 questões entre 1 e 34, terá raw score 21 e balanced score em torno de 80 (o escore exato dependerá de quais foram as 20 certas). Uma outra pessoa (B) que acertar todas as questões 1 até 25 e errar as outras, que são as mais difíceis, terá raw score 25, portanto maior que o escore de "A", mas seu balanced score será 10,3, isto é, muito menor que o obtido por "B". Até mesmo se a pessoa "A" tivesse acertado apenas a questão 35, mas errado todas as outras, seu escore seria 66, portanto muito maior que o escore de "B" (~10,3). Ao converter o raw score em QI, teríamos para A (21) o QI 151, e para B (25) teríamos o QI 162. Mas ao converter o balanced score em QI, encontraríamos para A o QI 175 e para B o QI 152, que seriam valores bem mais representativos das capacidades de A e B, sobretudo se considerarmos que apenas uma pessoa resolveu a questão 35 (seriam duas, contando com A) e 16 pessoas tiveram 100% de acerto nas 25 primeiras questões (seriam 17, contando com B). Isso é uma evidência de que são muito mais raras as pessoas que conseguem resolver a 35 do que as pessoas que marcam 100% nas 25 primeiras, ou que o nível de produção intelectual correspondente à solução da Q35 equivale ao trabalho conjunto de maior número de pessoas do que o nível requerido para resolver as 25 primeiras somadas. E mais: se "A" errasse todas as questões entre 1 e 34 e acertasse apenas a 35, seu QI seria 173, ou seja, até mesmo com 20 descuidos o seu QI cairia apenas 2 pontos. Mas com 20 descuidos o seu raw score equivaleria a um QI na faixa de 90 ou 95. A vantagem de usar escores ponderados por esse método, em vez de escores brutos, é imensa, podendo até mesmo permitir medições de QI com precisão de uma decimal e, o que é mais importante, com acurácia em torno de 1 ou 2 pontos até mesmo nas proximidades do teto do teste, porque cada setor do teste é normatizado com base nos escores combinados de todos os 123 testees. Então mesmo que o maior raw score de um testee seja 30, é possível projetar com exatidão quanto corresponde um raw score 34 ou 35. Isso é impossível de fazer pelos métodos convencionais de normatização, baseados em raridade, que acabam apelando para extrapolações mal fundamentadas e produzindo tetos completamente inverossímeis.

➢ Níveis de dificuldade relativos e absolutos:

O nível de dificuldade de cada questão depende do potencial médio do grupo e da quantidade de erros e acertos da questão, e pode ser calculado pela seguinte fórmula: **$Nd=pQI_m+k*ln(P)$.**

Nd é o nível de dificuldade do problema, expresso em QI.
pQI_m é o pQI médio do grupo de testees.
P é o peso da questão (Ne/Nc) no grupo.
k é a constante 11,16

O peso relativo é obtido por **P=Ne/Nc**. Se 1000 pessoas tentam resolver um problema "A", mas apenas 50 pessoas conseguem acertar, então este problema tem nível relativo de dificuldade 19 = [(1000-50)/50]. Se um outro grupo de 1000 pessoas tenta resolver um problema "B", mas apenas 5 pessoas conseguem acertar, então este problema tem nível relativo de dificuldade 199 = [(1000-5)/5]. Pergunta: Qual destes problemas é o mais difícil? Resposta: não dispomos de informações suficientes para responder. Nós precisamos conhecer o pQI médio do grupo que tentou resolver A e o pQI médio do grupo que tentou resolver B. Esta média pode ser

calculada de diferentes maneiras. A mais simples seria calcular a média aritmética dos QIs, mas conceitualmente isso não seria o método mais apropriado e só teria alguma validade porque a distribuição é aproximadamente gaussiana. Mas se o grupo não fosse gaussiano, então teríamos um erro imenso. Suponhamos, por exemplo, que uma única pessoa do grupo tivesse pQI 200 e as outras tivessem pQI 100, então a média aritmética do grupo seria 100,1. Mas a média do potencial para o mesmo grupo seria equivalente a um QI 124, porque a pessoa com QI 200 teria potencial de 7790 pessoas de QI 100, que somado ao potencial das outras 999 totalizaria quase 8789, então o potencial do grupo seria 8,8 vezes maior do que o potencial de um grupo com 1000 pessoas de QI 100 (todas com QI=100) e seria aproximadamente igual ao potencial de um grupo com 1000 pessoas de QI 124,3 (todas com QI 124,3). Então qual das médias representa melhor o potencial intelectual médio do grupo, 100,1 ou 124,3? Por tudo que vimos até aqui e veremos mais adiante, eu acredito que 124,3 é a melhor representação. Seria uma representação ainda melhor se fosse um grupo com 1.000.000 de pessoas, 1000 com QI 200 (todas as 1000 com QI 200) e 999.000 com QI 100 (todas as 999.000 com QI 100). Então este grupo teria o mesmo potencial de produção intelectual que um grupo com 1.000.000 de pessoas com QI 124,3 (todas com QI 124,3).

Mas um grupo de 1000 pessoas não terá uma distribuição tão exótica como nesse exemplo, então podemos usar como aproximação grosseira o pQI médio como sendo a mesma coisa que a média aritmética dos pQIs. Isso não causa grandes prejuízos ao método. Para os 123 testees de Sigma Test, por exemplo, a média aritmética dos QIs é 149,7 pela nova norma e 149,6 pela norma antiga. E o potencial médio de QI é 151,8 pelas normas nova e antiga. De modo geral, para distribuições gaussianas (ou similares), o potencial médio de QI será maior do que a média aritmética dos QIs quando o QI médio do grupo estiver acima de 100. Mas até mesmo para um grupo com QI médio 150, a diferença é de apenas 1 ou 2 pontos. Só haveriam diferenças marcantes, se a distribuição fosse forçada, como no exemplo anterior, e então poderíamos enxergar claramente que o método de média aritmética é errado.

Agora que temos um conceito apropriado para potencial médio de um grupo, podemos prosseguir. Como a diferença entre o potencial do grupo calculado pela média aritmética dos QIs não difere muito do calculado pelo potencial ponderado, para ilustrar esse argumento podemos usar indistintamente os dois métodos, sem prejuízos significativos. Contudo, no processo de normatização, é evidente que precisamos aplicar o método apropriado.

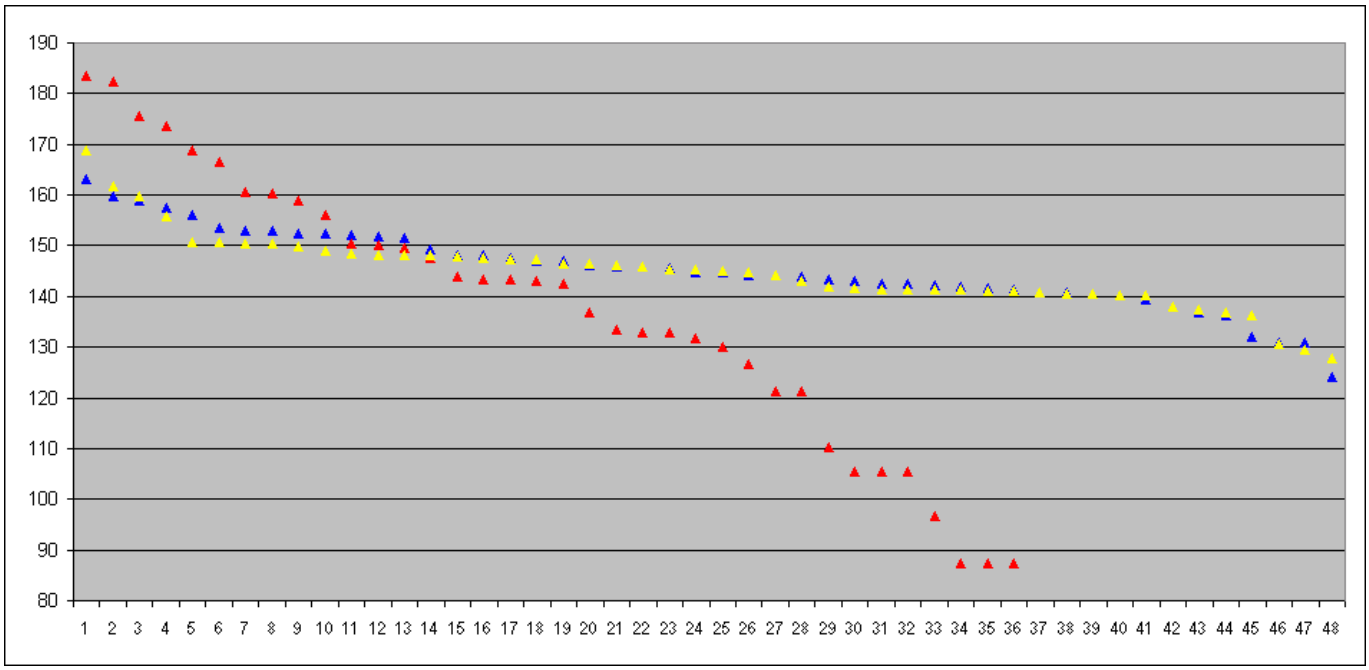
Retomando: se 1000 pessoas tentam resolver um problema “A”, mas apenas 50 pessoas conseguem acertar, então este problema tem nível relativo de dificuldade 19 = [(1000-50)/50]. Se um outro grupo de 1000 pessoas tenta resolver um problema “B”, mas apenas 5 pessoas conseguem acertar, então este problema tem nível relativo de dificuldade 199 = [(1000-5)/5]. Qual destes problemas é o mais difícil?

Resposta: isto depende do QI médio de cada grupo. Digamos que os dois grupos tenham mesmo QI médio, então é óbvio que a questão “B” é muito mais difícil do que a questão “A”. Mas se o primeiro grupo tem QI médio 150 e o segundo tem QI médio 100, então o nível de dificuldade da questão “A” é muito maior do que o nível de dificuldade da questão “B”. Por fim, se o primeiro grupo tem QI médio 130 e o segundo tem QI médio 100, então o nível de dificuldade das questões A e B é aproximadamente o mesmo.

Complemento em outubro de 2004: a afirmação acima vale se o parâmetro de discriminação a for o mesmo para os dois itens e dois grupos. Na época que desenvolvi o método, embora tenham surgido evidências de que as quantidades de acertos nos itens poderiam variar em proporções distintas com a variação do nível de habilidade, julguei que essas diferenças de itens individuais deveriam ser tratadas como flutuações estatísticas e não pensei na possibilidade de usar um parâmetro para descrever esse fato. A TRI tem um parâmetro para isso e nesse aspecto ela é superior ao método que propus em 2003.

Portanto, **P=Ne/Nc** fornece o nível relativo de dificuldade para um problema dentro de um teste. Mas o valor de P não diz nada quando são comparados testes diferentes. E **$Nd=pQI_m+k*ln(P)$** fornece o nível absoluto de dificuldade de um problema, e os valores de Nd podem ser comparados entre testes diferentes, mantendo uma proporção confiável entre os níveis de dificuldade.

Vejamos a seguir os níveis absolutos de dificuldade das questões do Sigma Test (vermelho), do Titan Test (amarelo) e do Mega Test (azul), em níveis de rarity-IQ, calculados usando pQI_m resultante do potencial de QI:



As questões estão dispostas em ordem decrescente de dificuldade (não é a mesma ordem que aparecem nos testes). Podemos observar que no Sigma Test as questões estão distribuídas em mais níveis de dificuldade, permitindo medir a inteligência numa gama muito mais ampla, enquanto o Mega e o Titan concentram-se no intervalo de 140-175 (135-155 no caso de rarity-IQ), assegurando maior precisão nesse patamar, mas não maior acurácia, devido ao fato de que estes testes não usam escores ponderados, originando erros como os que explicamos no tópico anterior, entre outros. O Titan e o Mega devem ter escores semelhantes na maior parte da norma.

Complemento em outubro de 2004: O comentário anterior, sobre o parâmetro de discriminação, pode ser ilustrado claramente nesse caso: o Sigma Test e mais discriminativo do que o Mega e o Titan, ou seja, variações nas porcentagens de acertos brutos indicam maiores variações no potencial. Isso acontece porque as questões mais fáceis do Sigma Test são mais fáceis do que as mais fáceis do Mega ou Titan, enquanto as mais difíceis do Sigma Test são mais difíceis do que as mais difíceis do Mega ou Titan. Isso torna a curva do Sigma Test mais inclinada. O gráfico acima é baseado na norma de setembro de 2003, mas não é muito diferente do que seria se fosse representado pela norma atual.

O mesmo cálculo usando pIQ_m obtido pela média aritmética dos pIQs produz um resultado semelhante, porém destacando as questões mais difíceis do Sigma Test e destacando.

➢ Como usar este método para determinar o QI:

Para cada pIQ existe um quociente de potencial (QP) que pode ser expresso da forma **$P_1/P_2=e^{[(pQI_1-pQI_2)/k]}$**, onde P_1 e P_2 representam os potenciais (para resolver problemas) das

pessoas 1 e 2, enquanto pQI_1 e pQI_2 são os respectivos pIQs das pessoas 1 e 2, e "k" é uma constante. Para o cálculo preliminar, foi utilizado para "k" o valor 10,03 (conforme calculado nesta página: http://planeta.terra.com.br/educacao/sigmasociety/medias_qi.html). Mais adiante, veremos que o melhor valor para k é 11,16 [O valor calculado com base na norma de outubro de 2004 é 10,82].

Conforme já vimos, o peso (P) de cada questão é determinado por P=Ne/Nc. Então o escore ponderado (EP) total de uma pessoa será dado pela soma dos pesos "P" de todas questões que essa pessoa acertar.

Partimos da hipótese (1) de que os escores ponderados (EP) correlatam fortemente com os QPs (quocientes de potencial) P_1/P_2. A justificativa para essa hipótese é que se entre 100 pessoas houver 5 que conseguem acertar a questão A, e se entre as mesmas 100 pessoas houver 5 que conseguem acertar as questões B, C, D, E, então podemos dizer que a dificuldade das questões B, C, D, E somadas é aproximadamente a mesma da questão A. Isso determina EP. Analogamente, se uma pessoa P_1 sozinha resolve as questões B, C, D, E, enquanto 10 outras pessoas P_n combinadas também resolvem as questões B, C, D, E (ou resolver a questão A, que é basicamente a mesma coisa), isso significa que a pessoa P_1 produz tanto quanto as 10 pessoas P_n, e se todas as pessoas P_n tiverem capacidades iguais, podemos dizer que P_1 é 10 vezes mais produtiva que cada uma das pessoas P_i. Isso determina P_1/P_2. Assim é fácil concluir que P_1/P_2 representa para as pessoas o mesmo que EP representa para as questões, então se uma pessoa A tem EP duas vezes maior do que uma pessoa B, podemos concluir que a pessoa A é duas vezes mais produtiva, ou seja: $P_A/P_B=2$.

Isso deve ser válido para um conjunto de perguntas com pesos determinados pela fórmula P=Ne/Nc e com o escore total ponderado determinado por $EP=\Sigma P_i$. Então os rIQs da norma antiga devem correlatar fortemente com os pIQs obtidos por uma função que tenha o seguinte aspecto: pIQ=100+k*ln(EP*q), onde "EP" é o escore ponderado, "k" é aproximadamente 11,16 e tem mesmo valor para todos os testes que pretendem fornecer escores com σ=16. Enquanto "q" é desconhecido e preliminarmente vamos calcular seu valor com base em nossa primeira estimativa para "k". Em meu primeiro cálculo, usei k=10,03. Para evitar distorções no teto e na base do teste, eliminei os 10% de escores mais altos e 10% mais baixos. Mesmo que houvesse tal distorção no teto e na base, ela seria dissolvida ao longo de toda a norma. Portanto a eliminação desses 10% é opcional, dependendo principalmente do tamanho da amostragem disponível. Eu também usei a média aritmética dos QIs, no primeiro cálculo, mas posteriormente melhorei isso usando o potencial do grupo, ganhando cerca de 2 pontos na acurácia. Para fins didáticos, será mantida a explicação usando média aritmética.

Complemento em outubro de 2004: a hipótese de que o valor de k é o mesmo para qualquer teste incorre no mesmo erro que cometi ao julgar que o parâmetro de discriminação é o mesmo para todo item. Aliás, k está para um teste assim como a está para um item, pois é k que determina o poder de discriminação do teste. Ao analisar os dados do Mega e Titan, constatei uma diferença marcante entre k para o Sigma Test e para os outros dois testes. Agora compreendo claramente esse fato, porque conforme vimos no gráfico acima, é natural que o parâmetro de discriminação seja diferente, sendo que a distribuição dos níveis de dificuldade das questões é claramente diferente. No entanto, seria esperado que o valor de k para o Mega e o Titan fosse mais semelhante entre eles do que entre o Sigma Test e qualquer deles, mas não é o que acontece.

➢ Como normatizar um teste usando este método:

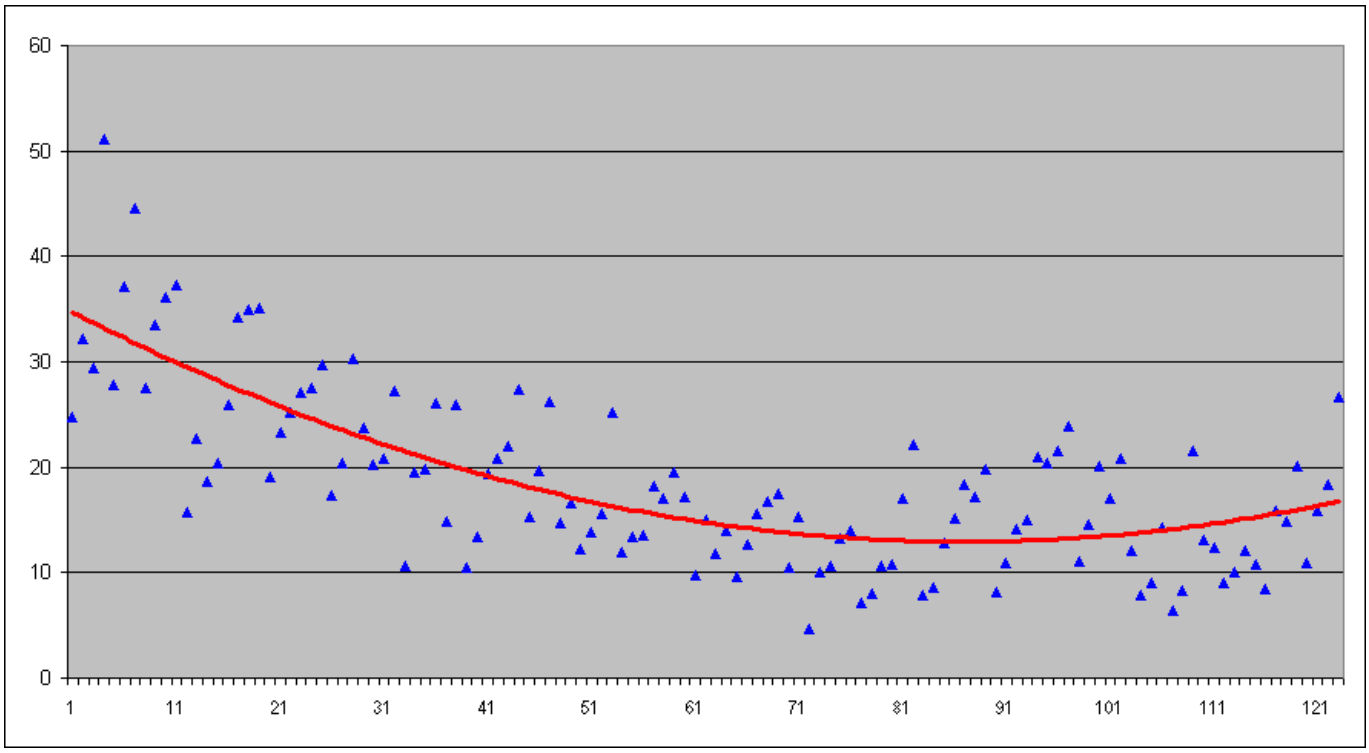
Antes de prosseguir, calcule os pesos (P) das questões de seu teste e calcule os escores ponderados (EP) para cada *testee* ("*testee*" é a pessoa que está sendo submetida ao teste, também pode ser chamada "*examinee*"). Depois siga estas instruções:

§1 – Use o valor preliminar k=10 e calcule o quociente de potencial (QP) para o QI de cada *testee* (use os QIs da norma antiga). $QP=e^{[(pQI_1-pQI_2)/k]}$, use $pQI_2=100$

§2 – Some os QPs (exceto os 10% mais baixos e os 10% mais altos). No Sigma Test obtivemos 23.191.

§3 – Some os EPs (exceto os dos QPs 10% mais baixos e dos QPs 10% mais altos). No Sigma Test obtivemos 1.263.

§4 – Divida o resultado da soma §2 pelo resultado da soma §3 e o resultado será o valor preliminar de "q". No Sigma Test obtivemos 18,365. Observe que se você dividir cada QP pelo respectivo EP, você terá um valor quase constante, a menos que os escores da norma antiga estiverem distorcidos. E foi justamente o que aconteceu nos casos do Sigma, Mega e Titan. O gráfico abaixo mostra os valores de cada QP divido pelo respectivo EP. O eixo y exibe os resultados de QP/EP e o eixo x mostra o ranking dos *testees* (1 é o QI mais alto etc.):



Embora a melhor regressão seja um polinômio de ordem 2, na nova norma a distribuição é forçadamente linear, porque mais adiante teremos que fazer um novo ajuste (§16 para §18) e seria irrelevante usar algum ajuste preliminar.

§5 – Aplique a fórmula $pIQ=100+k*ln(EP*q)$ e calcule o QI de cada testee para cada escore ponderado (EP).

§6 – Calcule o desvio-padrão dos QIs da norma antiga (exceto os 10% mais altos e os 10% mais baixos). No Sigma Test obtivemos 9,33.

§7 – Calcule o desvio-padrão dos pQIs recém obtidos (exceto os 10% mais altos e os 10% mais baixos). No Sigma Test obtivemos 7,75.

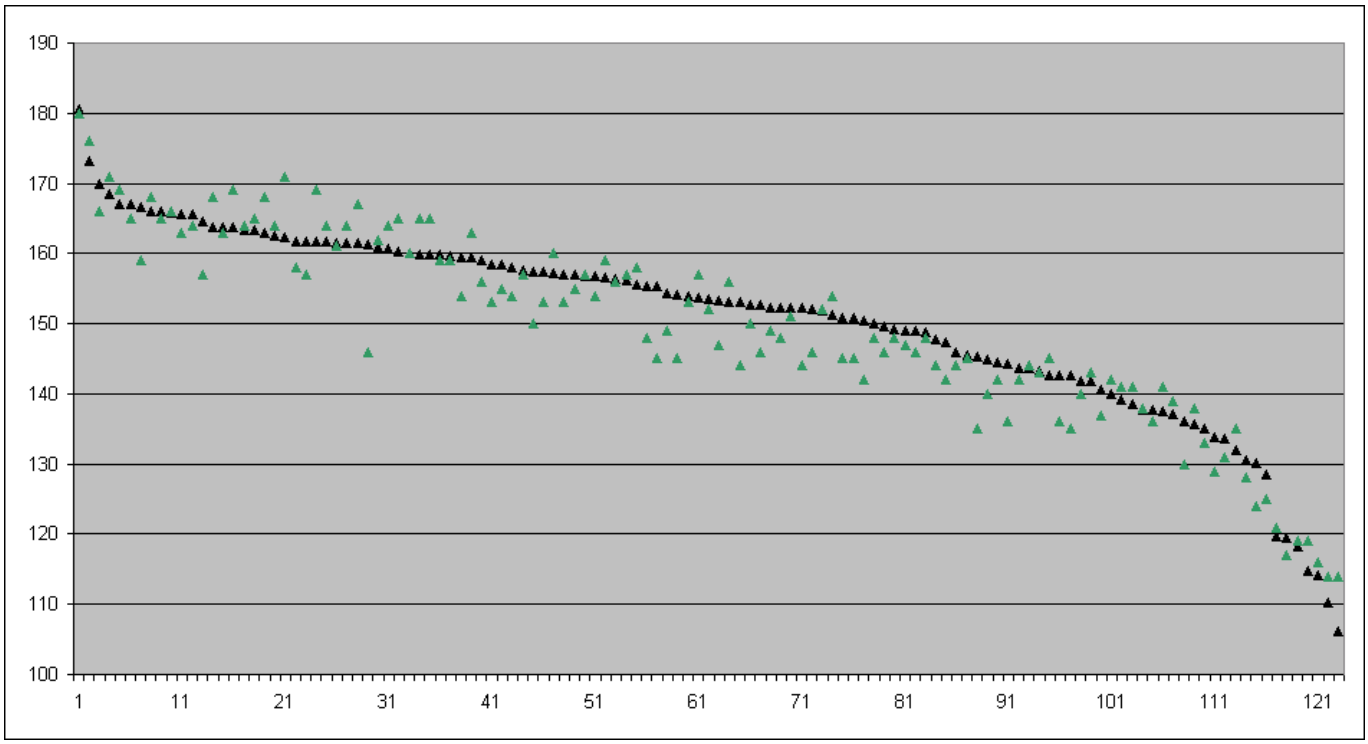
§8 – Altere o valor de "k" de modo que o desvio-padrão da norma antiga fique igual ao da norma nova. No caso do Sigma Teste, o valor encontrado foi 12,04. Para o Mega foi

12,75 e para o Titan foi 9,07. A média ponderada, com base nos escores de 891 testees, foi 11,16. Este valor precisa ser unificado para todos os testes. É recomendável que você use 11,16 se o seu teste não tiver várias centenas de testees, ou que você calcule um valor ponderado para "k" levando em conta os valores que você obteve combinados aos obtidos para o Meta, Titan e Sigma.

§9 – Ao mudar o valor de k, você obterá um novo valor para "q". Atualize o valor de "q".

§10 – Você pode colocar os QIs em ordem usando como referência os da norma antiga ou os da norma nova ou ambos (desfazendo os pares, e colocando todos em ordem decrescente ou crescente). A última alternativa parece ser a que produz normas mais acuradas, contudo, neste exemplo usaremos a primeira opção, por ser mais simples.

§11 – Munido de valores para "k" e "q" você já tem a pré-norma para pIQ. Na verdade, esta é uma norma de rIQ calculada com base no método de pIQ. A norma resultante é representada no gráfico pelos triângulos pretos, e os escores da norma antiga são os verdes.



A etapa seguinte consiste em calcular a norma de pIQ propriamente dita:

§12 – Disponha os rIQs em ordem crescente (ou decrescente) e os agrupe. Se tiver 50 testees, pode fazer 5 grupos de 10 testees ou 6 de 8. No Sigma Test, usamos 10 grupos de 12:

§13 – Calcule o QI médio de cada grupo (média aritmética dos rIQs). Se preferir, pode calcular o potencial médio.

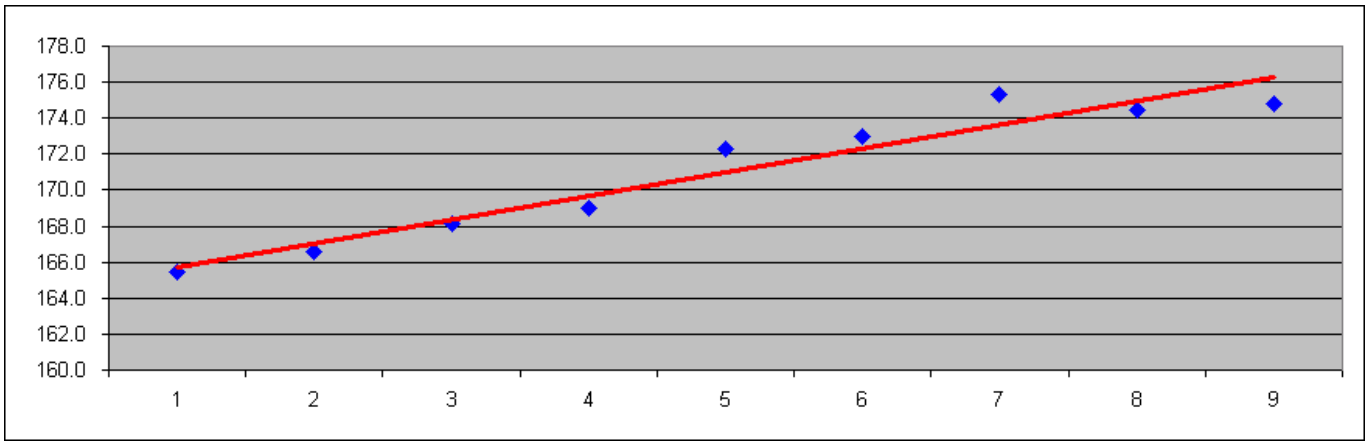
§14 – Calcule o nível de dificuldade de cada questão, com base no número de erros e acertos de cada grupo, usando a fórmula $Nd=QI_m+k*ln(P)$.

Nd é o nível de dificuldade do problema, expresso em QI.
QI_m é o QI médio de cada grupo.
P é o peso da questão (Ne/Nc) para cada grupo.

k é a constante 11,16

§15 – Calcule a média aritmética do nível de dificuldade de todas as questões para cada grupo de testees. Quando todos os testees de algum grupo tiverem acertado uma questão ou nenhum testee tiver acertado uma questão, você terá que deixar essa questão de lado. Não faça nada como calcular a dificuldade média das questões 1, 2, 3, 4 para o grupo de QI médio 135 e a média de dificuldade das questões 1, 3, 4, 5 para os demais grupos. Isso obviamente vai distorcer seus resultados.

§16 – Construa um gráfico com o nível médio de dificuldade de todas as questões calculadas para cada grupo de testees e observe se existe alguma tendência para o nível de dificuldade aumentar ou diminuir em função do QI médio do grupo. No caso do Sigma, do Mega e do Titan, foram observadas tendências para o nível de dificuldade calculado ser maior quando o QI médio do grupo é menor. Isso acontece porque o rIQ difere do pIQ. Se rIQ e p_IQ fossem iguais, nenhuma tendência desse gênero deveria ser observada. No Sigma Test obtivemos este gráfico:



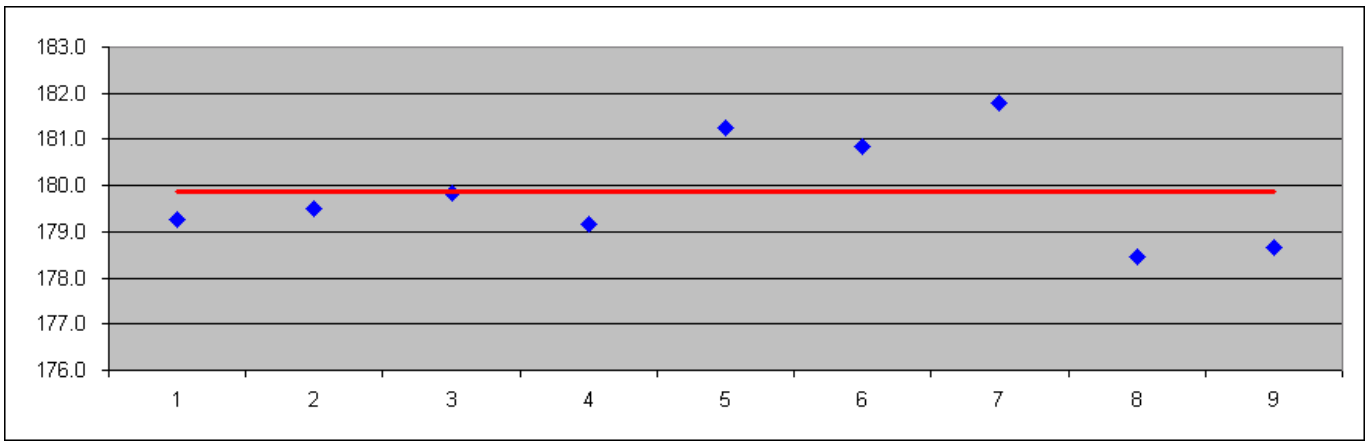
165.5	165.5
161.8	166.6
159.7	168.1
156.6	169.0
154.3	172.3
151.7	173.0
148.5	175.3
141.6	174.4
140.9	174.8

A tendência observada é notória: quanto menor o QI do grupo, maior é o peso calculado para as questões com base neste grupo. A tabela ao lado mostra os QIs médios dos grupos na coluna da esquerda e os níveis de dificuldade das questões na coluna da direita. O gráfico e a tabela foram obtidos usando os dados ordenados pela norma antiga. Usando a norma nova como referência, podemos observar uma tendência muito semelhante. A diferença está nos fatores de ajuste. Se for usada como referência a norma antiga, então os fatores de ajuste podem usar s=0,1444 na fórmula de Scoville ou usar σ=24,05 para pIQ. Nos dois casos nós teremos uma reta sem inclinação. Se a nova norma for usada como referência, então os fatores de ajuste podem ser s=0,1435 ou σ=25,20. A solução mais apropriada parece ser usar a fórmula de Scoville com s=0,1444.

Complemento em outubro de 2004: em lugar da fórmula de Scoville, na norma de outubro de 2004 usamos uma mais adequada para todo o espectro considerado e com fortíssima correlação com a fórmula antiga (maior que 0,99). Isso é descrito no texto sobre a nova norma: http://www.sigmasociety.com/artigos/norma_set_2004.pdf

§17 – Use a fórmula pIQ=100*e^[s*(rIQ-100)/16] e calcule os pIQs para cada rIQ (s=0,1444).

§18 – Repita os procedimentos §12 até §16, porém usando os escores recém calculados de pIQ, em vez de rIQ. Verá que a tendência será diferente da anterior. No Sigma Test obtivemos este gráfico:



§19 – Altere o valor de "s" de modo que a tendência não seja aumentar nem diminuir o nível de dificuldade calculado para cada grupo de pIQ.

Se a norma for baseada em poucos testees e o valor de "s" ficar maior que 0,17 ou menor que 0,11, é recomendável que você adote o valor unificado 0,1444 até que sua amostragem tenha aumentado. Feita esta calibração, verá que para quaisquer pIQs A_1, A_2, A_3, A_4... A_n, se a diferença de pIQ entre A_1 e A_2 for x e o QP entre A_1 e A_2 for y, então para todos os pares A_n, A_m cuja diferença de pIQ entre A_n e A_m seja x, o QP entre A_n e A_m também será y. Em outras palavras: João tem QI 120 e José tem QI 136. Isso significa que José produz 4,2 vezes mais do que João. Então se Pedro tem QI 150 e Eduardo tem QI 166, significa que Eduardo produz 4,2 vezes mais do que Pedro. E o mesmo se aplica a qualquer faixa de QI. Isso é válido para pIQ, mas apresenta ligeiras distorções se forem usados os rIQs. O motivo disso é que o pIQ é uma representação fiel do potencial intelectual, enquanto o rIQ é um sistema artifical para determinar o potencial intelectual com base na raridade, sujeito s todas as distorções causadas pela deterioração da gaussiana fora do intervalo -2σ a +2σ.

Alguns dados gerais sobre a nova norma do Sigma Test:

Correlação entre pIQ e rIQ: 0,958
Correlação entre rIQ da nova norma e da norma antiga: 0,952
Correlação entre raw score pelo novo método de correção e o antigo: 0,991
Correlação entre QPs da norma antiga e EPs da nova norma: 0,923
Desvio-padrão dos rIQs da nova norma: 13,50
Desvio-padrão dos rIQs da antiga norma: 13,98
Teto estimado: rIQ=199+, pIQ=244+ (Teto sem a questão 36: rIQ=193+, pIQ=232, Teto sem as questões 29 e 36: rIQ=185+, pIQ=216, Teto sem as questões 29, 35 e 36: rIQ=181+, pIQ=207, Teto sem as questões 29, 30, 35 e 36: rIQ=175+, pIQ=197)

A incerteza estimada no teto do Sigma Test é cerca de 0,4 ponto de QI (243,6 ±0,4). Um comentário muito interessante do nosso amigo Albert Frank merece ser incluído aqui e esclarecido. Ele diz aproximadamente isto:

"O nível de dificuldade da questão 35 do Sigma Test é 66, mas se amanhã uma pessoa fizer o teste e acertar esta questão, o nível cairá para 44. Portanto o teto do teste não pode ter uma incerteza de apenas 0,4."

O comentário é totalmente pertinente, mas se a pergunta 35 do ST tem peso 66 e amanhã uma pessoa a acertasse e o peso caísse para 44, isto afetaria o escore mais alto em 4 pontos (de 206 passaria a 202), e as outras 5 pessoas que tiveram fração de ponto por resposta

parcialmente certa, teriam variação de 0,2 em seus QIs. Os demais QIs não sofreriam nenhuma variação maior do que 0,05. Isso pode causar a ilusão de que os escores mais altos têm incerteza de 4 pontos e o teto teria incerteza ainda maior, mas não é o que acontece, porque seria preciso levar em conta a probabilidade de que o próximo *testee* acertará esta pergunta. Se a probabilidade fosse 1 em 2, então realmente haveria uma incerteza de cerca de 2 pontos no escore mais alto, mas todos os dados de que dispomos sugerem que a probabilidade é cerca de 1 em 67, então a incerteza é muito menor do que 4 pontos, provavelmente cerca de 0,06. Essa incerteza é baseada numa única pergunta. A incerteza combinada das 36 perguntas deve ser aproximadamente 0,06*35^½, ou seja: 0,36. Mas a maneira como esta incerteza é estimada pode ser inapropriada. Contudo, a incerteza provavelmente é pequena em praticamente toda a norma de 100 até 200 (erro menor que 1 ponto) e pode chegar a 5 pontos no teto. Isso é muito superior a qualquer teste que existe, cujo nível teórico de incerteza chega a 100 pontos até mesmo faltando 2 perguntas para chegar ao teto.

Complemento em outubro de 2004: pela nova norma a incerteza é calculada de maneira mais pertinente. Isso é descrito no texto sobre a nova norma: http://www.sigmasociety.com/artigos/norma_set_2004.pdf

Alguns dados gerais sobre a nova norma do Mega Test:

Correlação entre rIQ da nova norma e da norma antiga: 0,959
Correlação entre pIQ da nova norma e rIQ da norma antiga: 0,973
Correlação entre QPs da norma antiga e EPs da nova norma: 0,633
Desvio-padrão dos rIQs da nova norma: 17,87
Desvio-padrão dos rIQs da antiga norma: 20,53
Teto: rIQ=168,5+, pIQ=185,6+

Alguns dados gerais sobre a nova norma do Titan Test:

Correlação entre rIQ da nova norma e da norma antiga: 0,898
Correlação entre pIQ da nova norma e rIQ da norma antiga: 0,925
Correlação entre QPs da norma antiga e EPs da nova norma: 0,688
Desvio-padrão dos rIQs da nova norma: 15,70
Desvio-padrão dos rIQs da antiga norma: 13,43
Teto: rIQ=166,6+, pIQ=182,5+

Titan				Mega			
Raw	Norm	rIQ	pIQ	Raw	Norm	rIQ	pIQ
1	120	100.4	100.4	1	100	102.5	102.2
2	123	115.1	114.6	2	111	113.4	112.9
3	126	120.4	120.2	3	116	121.7	121.6
4	128	125.0	125.3	4	120	126.7	127.2
5	130	128.2	129.0	5	124	129.3	130.3
10	137	137.9	140.8	10	133	139.4	142.8
15	142	143.5	148.0	15	139	145.7	151.0
20	146	147.7	153.8	20	145	150.4	157.6
25	151	151.2	158.8	25	151	153.8	162.5
30	157	153.5	162.0	30	157	157.3	167.7
35	163	156.3	166.3	35	163	160.2	172.2
40	170	160.2	172.2	40	169	163.8	177.9
41	172	161.4	174.0	41	172	164.1	178.3
42	174	161.8	174.7	42	174	164.7	179.4
43	176	162.1	175.1	43	177	165.9	181.2
44	180	162.9	176.4	44	180	167.1	183.2
45	183	162.9	176.5	45	183	167.6	184.1
46	186	164.2	178.5	46	186	167.9	184.5
47	190	165.4	180.5	47	190	168.1	185.0
48	190+	166.6	182.5	48	190+	168.5	185.6

A tabela à esquerda mostra as novas normas para o Mega e Titan, com raw scores, QIs pela norma antiga, rarity-IQs pela nova norma e potential-IQs pela nova norma.

Todos os testes considerados apresentaram melhor correlação entre pIQ da nova norma com rIQ da norma antiga do que entre rIQ das normas nova e antiga. Além disso, a média aritmética dos pIQs é mais semelhante ao QI de potencial médio dos testees do que a média aritmética dos rIQs. Isso sugere que o pIQ seja uma melhor representação do QI do que o rIQ, não apenas por apresentar melhor correlação com os testes Mega, Titan e Sigma, mas principalmente por não apresentar tendência de variação nos QPs para diferentes faixas de pIQs, como acontece no caso dos rIQs.

A calibragem da nova norma foi feita com base no intervalo de confiança da norma antiga (cortando os 10% maiores escores e os 10% menores escores). Existem outras maneiras de calibrar a nova norma, como usando pontos específicos seguros. Digamos que um teste teve 200 testees com escores entre 1350 e 1400 no SAT, e 600 testees com escores 1100 e 1150, então isto pode ser usado como referência. Se houver 3 ou mais pontos de referência, é ainda melhor. Se houver uma larga faixa, em vez de intervalos estreitos, pode ser ainda melhor.

Ainda existem problemas na norma do Titan e do Mega, um dos quais é a pergunta da fita de Moebius, que pode ser resolvida usando um método certo e muitos métodos errados. Para desgraça da norma do teste, se a pessoa usar um dos métodos errados e ridiculamente fácil, ela encontrará a mesma resposta do método certo. :-(Isto estraga a norma inteira, incutindo uma incerteza intrínseca de 1 a 5 pontos em todos os escores (as incertezas são maiores nos QIs mais baixos). O reflexo disto é que no Titan surgem alguns resultados absurdos, como uma pessoa que acertou só 1/48, e este único ponto certo foi a pergunta da fita de Moebius. :-\ Como não é possível saber se a pessoa usou o método certo ou não, eu pensei em calcular a

norma novamente, e excluir esta questão. Mas eu também desconfio que existe alguma solução ridícula que fornece resultado correto na questão dos cones/cilindro interprenetration. Contudo, estas perguntas são excelentes e muito importantes para manter o teto com nível alto. A melhor solução seria se Hoeflin pedisse um esboço da resolução, não apenas a resposta numérica.

Outras questões do Mega e do Titan também padecem deste problema, porque o número de respostas "não-estúpidas" oscila num intervalo pequeno. Se a pergunta pode receber respostas "não-estúpidas" entre 6 e 15, é como se fosse uma pergunta de múltipla escolha com 10 alternativas, e o perigo de uma pessoa acertar acidentalmente é grande. O efeito que isto produz é puxar o teto para baixo e puxar o resto dos escores para cima! Isto explica pelo menos uma parte da distorção surpreendente que nós vemos no cálculo acima. O nível de dificuldade das questões mais difíceis talvez seja 10 pontos maior, e o teto “verdadeiro” pode ser realmente 10 pontos maior, contudo a única maneira de calcular objetivamente o teto é esta, e o resultado é ~168.

➢ Algumas críticas:

A relação de proporcionalidade QP se aplica em toda a distribuição de QIs, fornecendo proporções acuradas para QIs entre -∞ e +∞. Contudo as limitações do teste só fornecem escores que podem ser convertidos em QPs no intervalo entre 100 e 200. Para QIs muito acima de 200 ou abaixo de 100, a acurácia é menor. Uma formiga, portanto, pode ter QI -3220 enquanto um alienígena pode ter QI 755. Então seriam necessárias 10^129 formigas para produzir intelectualmente tanto quanto uma pessoa normal, ou seja, impossível em nosso universo, porque uma quantidade tão grande de formigas seria muito maior que a quantidade de átomos no universo (10^79). E a quantidade de pessoas necessárias para igualar o nível de produção intelectual de um ET com QI 755 seria 10^25, portanto muito maior que o número de todas as pessoas que já viveram (10^11). Um comentário interessante do nosso amigo Albert é que não importa a quantidade de formigas somadas, elas nunca poderiam igualar a capacidade de produção intelectual de uma pessoa normal, mesmo 10^1000 formigas. Infelizmente não temos como saber isso, porque o máximo que conseguimos pensar em alguns milhões de formigas, e alguns milhões realmente não poderiam produzir tanto quanto uma pessoa normal, mas já não conseguimos imaginar bilhões ou trilhões, e se não conseguimos pensar nem mesmo em 10^12 formigas, como poderíamos saber com segurança de uma quantidade cem mil googols de trilhões vezes maior? Simplesmente não podemos saber. Mas podemos mudar progressivamente o patamar pata cada nível de comparação, e em vez de pensar em 10^129 formigas, podemos pensar que 100 chimpanzés produzem intelectualmente tanto quando 1 pessoa normal, e 1000 porcos produzem tanto quanto um chimpanzé, 1000 ratos produzem tanto quanto um porco e assim sucessivamente, até chegar nas formigas. É basicamente o mesmo que acontece no Xadrez. Kasparov, com seus 2850 de rating, teoricamente venceria 99,998% dos jogos contra uma criança com rating 1000. Então eles precisariam jogar pelo menos 1.000.000 de jogos para que a criança marcasse uns 20 pontos e a tese fosse mais ou menos confirmada. Se eles jogassem “apenas” 10.000 jogos (1 jogo por dia durante 27 anos), talvez a criança perdesse 100%. Se eles jogassem o dobro ou o triplo, talvez a criança ainda perdesse 100%. Se jogassem 100.000 jogos, ainda assim talvez a criança perdesse 100%. Mesmo que jogassem 1.000.000 jogos, talvez a criança perdesse todas. Por outro lado, seria possível que a criança marcasse 1 ponto em menos de 1000 jogos. Mas para ter uma amostragem mais ou menos segura, seria preciso que jogassem no mínimo alguns milhões de jogos, o que é inviável. Por outro lado, a criança poderia jogar com seus colegas de 1400 e marcar 9%, e estes jogariam com os de 1800 e marcariam 9%, que por sua vez jogariam com os de 2200 e marcariam 9% etc. Com base nisso, podemos presumir o escore provável de um confronto entre o Kasparov e o menino com rating 1000. Da mesma maneira, para medir a

distância dos quasares nós começamos medindo uma vareta e uma sombra, depois medimos o tamanho de uma fração de um meridiano, em seguida calculamos o tamanho do planeta, depois usamos a paralaxe causada pelo raio do planeta para calcular a distância do Sol, então usamos a órbita em redor do sol para produzir paralaxes estelares e calcular as distâncias de estrelas próximas, depois usamos vários recursos combinados (paralaxe geométrica, paralaxe estatística etc.) para calcular as distâncias de cefeídas em nossa galáxia, então usamos os brilho / período das cefeídas para calcular as distâncias de galáxias próximas, depois usamos o desvio para o vermelho e conseguimos calcular a distância dos quasares. São vários elos que ligam uma ordem de grandeza à outra, intersectando dois métodos distintos. No último elo, nós usamos cefeídas e red-shift, e prolongamos o red-shift até as fronteiras do universo. É verdade que em todas essas etapas nós estamos propagando incertezas, mas é melhor conhecer uma grandeza com um certo grau de incerteza do que não conhecer nada. Se nosso método for adequado para calcular o QI de uma esponja e de um alienígena 100 milhões de anos mais evoluído do que nós, ainda que a incerteza na medida seja de 20%, será um avanço extraordinário em comparação ao método atual, que nem sequer admite QIs abaixo de 0 ou acima de 500.

A idéia de proporção entre os níveis de produção intelectual é um critério encadeado, que permite comparar pares de entidades conscientes com níveis semelhantes e, a partir daí, calcular escores que representem comparações entre pares com níveis muito diferentes. Pessoas com QI 150 podem ser facilmente e precisamente comparadas a pessoas com QI 100 ou QI 200, mas pessoas com QI 100 dificilmente podem ser comparadas a pessoas com QI 200. Uma pessoa com QI 150 produz tanto quanto 88 pessoas com QI 100 e uma pessoa com QI 200 produz tanto quando 88 pessoas com QI 150, e disso podemos inferir que uma pessoa com QI 200 produz tanto quando 7750 pessoas de QI 100. O problema é que uma quantidade tão grande quanto 7750 dificilmente conseguiria trabalhar organizadamente em conjunto, dificultando a validade da hipótese seja verificada. A idéia de proporção vale para QIs entre $-\infty$ e $+\infty$ apenas em teoria. Na prática, podemos constatar que as proporções funcionam bem para QIs entre 120 e 200, e podemos concluir que sempre funcionam para diferenças menores que 80 pontos, sendo especialmente eficientes para intervalos menores que 40 ou 50 pontos. A suposição de que funcionam para intervalos mais dilatados é razoável, mas ainda não temos como comprovar isso.

➢ Algumas das vantagens desse método de normatização:

1 – A principal vantagem é conceitual. Os primeiros escores de QI eram números com significado nebuloso que mudavam em função da idade. Uma criança de 2 anos com QI 300 passava a ter QI 250 aos 4 anos, QI 200 aos 10 anos e QI 160 na idade adulta. Teoricamente o escore deveria representar a proporção entre idade mental e cronológica, mas essa hipótese era ruim, porque a proporção não se mantém constante. Depois disso, começaram a usar a raridade para determinar os escores de QI. Mas este também é um método defeituoso. Suponhamos duas pessoas fortuitas de uma população com média de altura 1,7m e σ=7cm. Digamos que por sorte (ou azar) estas pessoas medem 0,7m e 2,7m. Então, como nossa amostra contém dois elementos, podemos dizer que a raridade para 2,7m é 1 em 2 com incerteza de 500.000.000.000.000.000.000.000.000.000.000.000.000.000.000 (500 milhões de trilhões de trilhões de trilhões), porque a distribuição teórica prevê que apenas 1 que em cada 5*10^44 pessoas podem ter altura 14,29 desvios padrão acima ou abaixo da média. Isto obviamente não faz nenhum sentido. Nem tampouco faria sentido dizer que estas pessoas devem medir 1,7m com incerteza de 1m. A maneira razoável de determinar as alturas dessas pessoas e calcular as incertezas nessas alturas é usando uma régua, e dizer que elas medem 0,7m e 2,7m com incerteza 0,005m, com base no limite de precisão da régua, ou então medir as pessoas várias vezes, calcular a média, o desvio-padrão e o desvio-padrão na média. Esse

segundo método deveria resultar em algo perto de 0,001m. As grandezas estimadas com base nos níveis de raridade não fazem sentido, porque dependem das particularidades da população em que as medições são feitas e de uma infinidade de fatores que degringolam a distribuição fora dos limites de 2 desvios-padrão distantes da média. Além disso, conforme vimos anteriormente, os testes cronometrados apresentam muitas deficiências intrínsecas, que os impede de medir a inteligência acima do QI 140. Então podemos concluir que o procedimento conceitualmente correto e operacionalmente acurado para medir a inteligência consiste em calcular proporções entre níveis de produção intelectual. E o Sigma Test é o primeiro instrumento usado com esse propósito. Agora, pela primeira vez na história, é possível medir a inteligência por uma escala não-arbitrária e não-distorcida, cujos escores representam logaritmos de proporções representativas das verdadeiras capacidades intelectuais das pessoas. E isso pode ser vastamente confirmado na vida real. Por exemplo: se um grupo de 10 alunos com QIs entre 129 e 131, trabalhando juntos, conseguem uma nota 9,5 num determinado trabalho escolar, então é esperado que um grupo com 147 alunos com QIs entre 99 e 101, também trabalhando juntos, conseguirão produzir um trabalho escolar com mesmo nível de qualidade e merecedor de uma nota 9,5. Outro exemplo: um grupo de 10 pessoas com QI 130 deseja construir um relógio analógico, mas nenhuma delas sabe como fazer isso e não dispõem de um manual de instruções. Dispõem apenas das peças. Digamos que este grupo consiga resolver o problema em 10 dias. Então é esperado que 147 pessoas com QI 100 também consigam resolver o mesmo problema em aproximadamente 10 dias. O mesmo acontece a praticamente todos os problemas que exigem pensamento. Usando amostras de 1140 pessoas dos testes Mega, Titan e Sigma, pudemos confirmar esta hipótese em praticamente todos os níveis de QI entre 110 e 200, e presumivelmente a hipótese continua a ser válida fora deste intervalo, talvez com diferenças muito pequenas, que podem ser ajustadas à medida que nossa amostragem se tornar maior.

2 – A acurácia no teto do Sigma Test é muito superior à do teto de qualquer outro teste. Isso acontece porque a medida da grandeza propriamente dita assegura uma precisão genuinamente muito superior. Outros testes calculam a incerteza em termos de raridade, o que em si constitui um erro, porque em níveis de raridade não acontece propriamente uma "incerteza". O que acontece é uma "inadequação". Se temos um grupo de 1000 pessoas com altura média 1,7m, σ=0,07m e a pessoa mais alta tem 1,95m. E outro grupo de 1000 pessoas com altura média 1,7m, σ=0,07m e a pessoa mais alta tem 2,70m. Se medirmos a grandeza "altura" (usando régua) e depois converter isto em "raridade", teremos resultados muito diferentes nos dois grupos, sugerindo uma incerteza de muitos sextilhões de sentilhões no nível de raridade, quando na verdade a incerteza na grandeza é de apenas alguns milímetros. Podemos ter uma acurácia muito boa na medida da altura (0,001m), e esta acurácia é verdadeira, baseada na grandeza que estamos medindo. Mas quando tentamos converter isto numa coisa que não correlata bem com a grandeza, ou seja, na raridade, então o que encontramos é um número "inadequado". A única utilidade da raridade pode ser estimar a probabilidade de encontrar pessoas com determinado QI numa população hipoteticamente gaussiana. No mundo real, em que as gaussianas degrigonlam, ainda pode ser tolerável falar em “raridade” para intervalos menores que 1,5σ a partir da média, ou 2,5σ quando as amostras incluírem vários milhões de testees.

3 – O QI não precisa ser obtido por uma tabela. Em vez disso, o QI sai diretamente de uma fórmula, que correlata 0,958 com os QIs obtidos pela norma antiga. Isso evita problemas de interpolações e proporciona resultados muito mais acurados, inclusive permitindo representar os QIs com uma decimal de precisão.

4 – Os pesos das questões são baseados em estatísticas, em vez de seguir uma escala arbitrária (ou, pior ainda, serem considerados todos iguais), melhorando ainda mais a precisão, porque uma pessoa que errar 5 ou 10 questões fáceis por descuido, poderá não perder nem

sequer 1 ponto no QI, ao contrário do que acontece nos testes com pesos iguais. A ponderação das questões assume uma importância tão grande que se uma pessoa que não conhece nada sobre Matemática quiser resolver exclusivamente as questões 34 e 35, ela pode ter escore 176, enquanto outra pessoa que resolva todas as questões matemáticas (1-25, 31 e 32) terá escore 162.

É importante esclarecer que existem muitas maneiras ruins de determinar os pesos das questões. Por exemplo: se em vez de P=Ne/Nc a fórmula fosse P=T+Ne-Nc (onde T=total de testees), ou P=100*Ne/T, haveria grandes distorções nos extremos superior e inferior. A forma mais adequada é P=Ne/Nc.

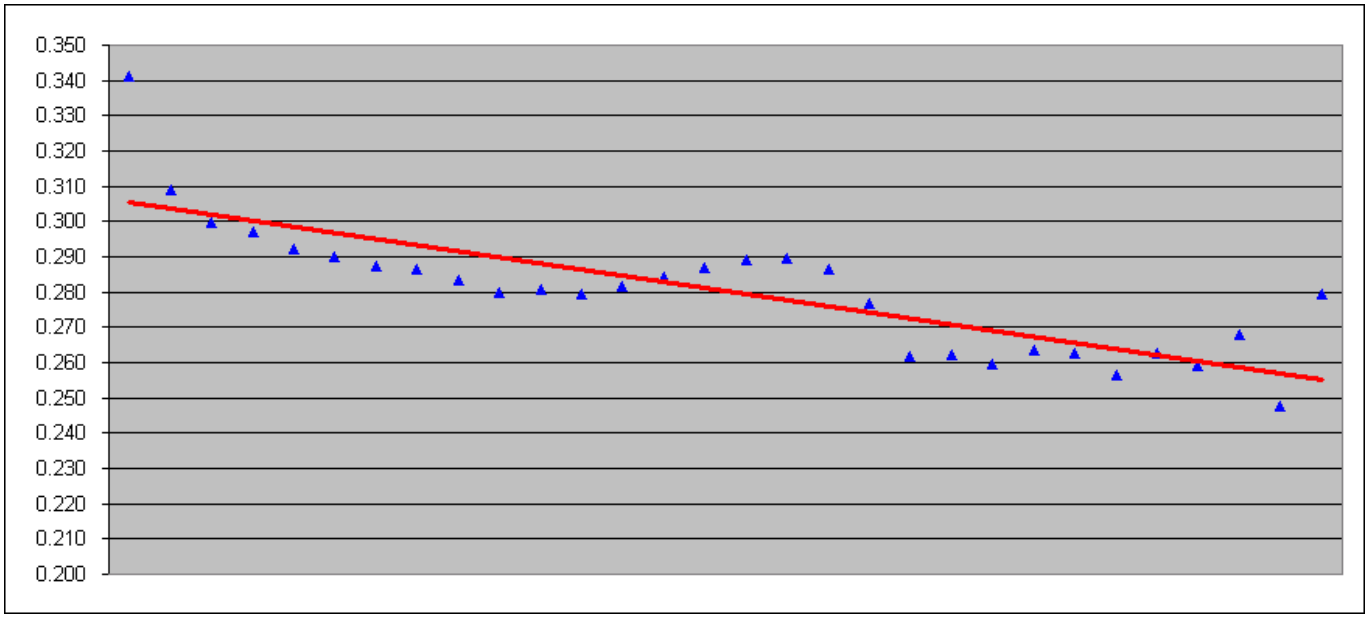
5 – O teto de um teste pode ser determinado com segurança, mesmo que nenhum *testee* tenha marcado perto de 100%. Por exemplo: mesmo cortando os 10% de escores mais altos do Mega e Titan, que incluem todos os raw scores acima de 40 no Mega e todos acima de 32 no Titan, encontramos os tetos 168 e 167, respectivamente. Nenhum outro método existente permitiria calcular tetos de 48/48 com base em escores abaixo de 33/48. E embora não haja uma maneira clara para calcular a incerteza nesse teto, podemos comparar ao que sucederia se não fossem cortados esses 10% mais altos e mais baixos, e fossem considerados todos os escores. Em tal caso, os tetos calculados seriam 171 e 172, respectivamente. As diferenças de 3 e 5 pontos sugerem inflação na norma antiga, causada pela incerteza no teto combinada à "motivação" para tornar o teto o mais alto possível.

6 – Conhecendo o nível de dificuldade de uma questão, é possível usá-la para normatizar outro teste. No caso do Sigma Test VI, por exemplo, a presença das questões 1 e 2, que também estão no Sigma Test, possibilita estabelecer uma norma preliminar satisfatória até o nível aproximado de 175. A incerteza é grande, contudo a confiabilidade é comparativamente maior do que em alguns testes que usam normas baseadas em amostras grandes e meticulosamente calculadas pelos métodos tradicionais, mas não usam questões com níveis apropriados de dificuldade, de modo que eles realmente conseguem medir com mais precisão, mas não se sabe exatamente o que estão medindo.

O método foi especialmente desenvolvido para o Sigma Test, que usa questões com pesos diferenciados e até mesmo respostas com diferentes pesos para a mesma questão. Mas também pode ser usado com sucesso em outros testes. O principal diferencial está em usar um sistema objetivo de balanceamento, por meio do qual é possível não apenas ordenar as questões pelo nível de dificuldade, mas também dizer quantas vezes uma questão é mais difícil do que outra, e com base no mesmo princípio, produzir escores que permitam dizer quantas vezes uma pessoa é mais produtiva do que outra. Depois de tudo que vimos, agora sabemos que uma pessoa com QI 118 é 5 vezes mais produtiva do que outra pessoa com QI 100. Uma pessoa com QI 136 é 25 vezes (5^2) mais produtiva que uma pessoa com QI 100. Uma pessoa com QI 172 é 625 vezes (5^4) mais produtiva que uma pessoa com QI 100. E assim por diante. E o próprio escore ponderado já fornece uma boa relação desse gênero: quem marca 20 pontos é 4 vezes mais produtivo do que quem marca 5 pontos.

Usando este método, só precisamos de uma média aritmética para os QIs de um grupo distribuído aproximadamente em gaussiana, e devolvemos uma norma inteira baseada nisto, sem distorções grandes no teto e na base. A norma inteira usa todos os testees da amostragem, assim a acurácia no teto é quase tão boa quanto a acurácia na média. Para o Mega, por exemplo, eu posso dizer que o teto é 168,51 e tem incerteza 0.25, porque das 32 pessoas que acertaram a pergunta mais dura, o QI médio é 164,25 com σ=1,90. Se eu calcular o desvio-padrão na média, terei 0,341. Se eu omitir o escore mais baixo que acertou a pergunta mais dura, e fizer o cálculo usando os 31 escores restantes, eu terei σ=1,69 e erro na média=0,309. Eu posso repetir o procedimento para 30, 29, 28 etc., e depois montar um gráfico

com estes valores para os top 3 a top 32 escores (exclusivamente os escores de quem acertou a pergunta mais dura). Neste caso nós temos uma reta, conforme podemos ver abaixo, e a linha de tendência parece convergir para 0.25 quando n=0. Se eu quiser uma estimava menos elaborada para a incerteza, eu posso fazer apenas o método com 32 escores e dizer que a incerteza é menor que 0,34:



Esperamos que esse processo de normatização rapidamente substitua a primária metodologia vigente, que usa escores não-ponderados e calcula as normas com base na falsa hipótese de a as distribuições são gaussianas em toda a extensão.

Complemento em outubro de 2004: A TRI consegue resolver vários problemas descritos aqui, de modo que algumas das vantagens que atribuí inicialmente ao meu método só representam vantagens em comparação à TCT (Teoria Clássica dos Testes), mas não à TRI. Os pontos ponderados constituem uma vantagem inclusive em relação à TRI, e o conceito de pIQ também. Mas os métodos de padronização da TRI, bem como o cálculo das incertezas, não é tão bom como o usado em TRI. Na norma de 2004 vários desses detalhes foram aprimorados e agora temos mais segurança em afirmar que se trata do melhor método para padronização de testes, tanto conceitualmente quanto tecnicamente.

Hindemburg Melão Jr., 23 de agosto de 2003

Dados do Mega Test: http://www.eskimo.com/~miyaguch/megadata/item_ana.html
Dados do Titan Test: http://www.eskimo.com/~miyaguch/titandata/item_ana.html
Artigo de Grady Towers: http://www.eskimo.com/~miyaguch/grady/about_rasch.html
Artigo de Bill McGaugh: http://www.pe.net/~bmcgaugh/eloiq.htm