Earnings inequality and skill mismatch in the U.S.: 1973-2002

Slonimczyk, Fabian

Higher School of Economics

2011

Earnings Inequality and Skill Mismatch in the U.S: 1973–2002*

Fabián Slonimczyk†

November 7, 2011

Abstract

This paper shows that skill mismatch is a significant source of inequality in real earnings in the U.S. and that a substantial fraction of the increase in wage dispersion during the period 1973–2002 was due to the increase in mismatch rates and mismatch premia. In 2000–2002 surplus and deficit qualifications taken together accounted for 4.3 and 4.6 percent of the variance of log earnings, or around 15 percent of the total explained variance. The dramatic increase in over-education rates and premia accounts for around 20 and 48 percent of the increase in the Gini coefficient during the 30 years under analysis for males and females respectively. The surplus qualification factor is important in understanding why earnings inequality polarized in the last decades.

JEL classification: J31

Key words: Skill Mismatch, Earnings Inequality, Shapley Value Decomposition

*I would like to thank Peter Skott, seminar participants at Amherst, Moscow, Mannheim, and Maastricht, and two anonymous referees for their comments and suggestions. All remaining errors are my responsibility.

†International College of Economics and Finance. Higher School of Economics, 11 Pokrovsky Bulvar. Moscow, Russian Federation. fslonimczyk@hse.ru.

1 Introduction

Over the last three decades the U.S. wage structure has widened. While the general trend has been toward greater inequality of earnings, the evolution of the wage distribution has followed a complex pattern.[1] The literature that was spurred has investigated the degree to which the rise in income and wage inequality can be attributed to changes in the returns to skill (eg. Katz and Murphy, 1992; Murphy and Welch, 1992; Bound and Johnson, 1992)[2]. There are two good reasons to focus on skill differentials. First, in decomposition exercises inequality between skill groups accounts for around a third of the variance of earnings in a cross-section of the working population. Since no other observable factor comes close in explaining variation in earnings at a point in time, it is natural to expect that changes in the wage distribution largely follow changes in the distribution of skills and their prices.

Second, skill differentials are easily interpreted in the light of a competitive model of the labor market. If different skill groups are imperfectly substitutable inputs in production, then shifts in their relative supply and demand curves can explain changes in their relative prices. For example, it is widely understood that during much of the 1970s the college premium decreased substantially because the well-educated baby boom generation entered the labor market increasing the relative supply of young skilled workers (Freeman, 1976). In the decade that followed, despite the continuous rise in the relative supply of skilled workers, there was a very significant increase in the wage differentials by education and labor market experience. Thus, several

[1] For example, a well-known exception to increasing earnings inequality is the shrinking male–female wage gap (Blau and Kahn, 2006). The analysis in this paper is conducted for males and females separately.

[2] See also the reviews of the literature in Levy and Murnane (1992) and Katz and Autor (1999). DiNardo et al. (1995) and Lee (1999) made the case that other important forces —changes in labor market institutions in particular— are fundamental in understanding the changes in the wage structure since the 1970s.

studies concluded that during the 1980s there was a strong shift in the relative demand for high-skill workers. Skilled biased technological change [SBTC] —the computer revolution in particular— was the leading candidate cause for the relative demand shifts.[3]

Despite the justified focus on skill wage differentials, it is a fact that residual or within-group wage inequality —i.e wage dispersion among workers with the same education and experience— accounts for the majority of the increase in the variance of earnings.[4] One possible explanation for the increase in residual inequality involves unobservable differences in human capital. If individuals differ in innate ability levels, schooling quality, motivation, etc., then an increase in either the dispersion of the unobservable abilities or the rewards that accrue to them could account for the rise of inequality within groups.[5] Another possibility is that within-group inequality is due to individuals' "behavioral traits" which are not productive skills (Bowles et al., 2001). Because the groups that happen to have greater within dispersion —highly educated and more experienced— have become more prevalent over time, the increase in residual inequality is partly a mechanical outcome of these compositional changes (Lemieux, 2006a).

In this paper, I explore an alternative story that relies on the dispersion of outcomes within education groups because of the existence of skill mismatch. According to assignment models of the labor market, not all workers are allocated to jobs in which their skills are required.[6] Some workers will be over-qualified, meaning that the skills

[3]Among others, this point was raised by Bound and Johnson (1992); Juhn et al. (1993) and Acemoglu (2002). A critical view is articulated in Card and DiNardo (2002).

[4]According to Katz and Autor (1999, p. 1490), 60% of the increase in the variance of log weekly wages over the 1963–1995 period was due to the growth of residual inequality.

[5]For example, using the assumption that all residual inequality is due to unobserved skill differences, Juhn et al. (1993, p. 429) find that about two-thirds of the increase in the 90–10 percentile gap of log wages over the period 1964–88 is due to changes in unmeasured prices and quantities.

[6]Assignment models are reviewed in Sattinger (1993). The task-based model in

they possess are above those required on the job. Similarly, some workers might have less qualifications than those required. Normally, over(under)-educated workers will have lower(higher) earnings than correctly matched workers with the same levels of skill. An increase in these match differentials or in the prevalence of mismatch has the potential to explain the increase in residual dispersion that accounts for the majority of the growth in earnings inequality. While there are many studies that look at match differentials,[7] to my knowledge this is the first paper explicitly linking the growth in earnings inequality to over- and under-qualifications.

The idea that skill mismatch might contribute to explain the changes in the wage structure is consistent with the nuanced version of the SBTC hypothesis as developed in Autor, Levy, and Murnane (2003) [ALM] and Goos and Manning (2007). According to ALM, technological change generally works toward substituting routine cognitive and manual tasks but is complimentary to non-routine tasks. Routine tasks are prevalent in "middling jobs", i.e. jobs that require average or just above-average skills and have traditionally paid around median wages. The ALM hypothesis predicts technological change reduces relative demand for these jobs and increases relative demand for jobs intensive in non-routing tasks. On the one hand, the requirement of non-routine skills is what distinguishes the best jobs in the market: managerial and professional positions. On the other hand, the worst paid jobs in the economy –eg. cleaning– are also heavy in non-routine tasks that cannot be substituted by technology. All in all, ALM provide a compelling argument that technology leads to increasing relative demand at these two extremes of the job distribution and decreasing demand for middling jobs. If this process of job

Acemoglu and Autor (2011) also has the property that skill groups can be assigned to a range of tasks of varying complexities. A skill mismatch equilibrium is also present in the search model in Albrecht and Vroman (2002) and the efficiency wage model in Skott (2006).

[7] I present a brief review of this literature in section two. For a comprehensive survey see McGuinness (2006).

polarization is accompanied by a continuous increase in the educational qualifications of the labor force, then it is to be expected that more and more highly educated individuals will end up in positions for which they are over-qualified. In this paper I attempt to determine to what extent this story is supported by the data.

Following the modern literature on skill mismatch, I start by specifying an extended earnings function that decomposes education into three parts: required, surplus and deficit qualifications. Using the Shapley value decomposition method developed in Shorrocks (1999), I then show that considering skill mismatch factors significantly increases the fraction of earnings inequality that can be explained at any point in time. I analyze not only the variance of earnings but also the Gini coefficient, and the 90–10, 90–50 and 50–10 percentile gaps.

While the added explanatory power of the extended equation is modest, the decomposition of the changes in the inequality indexes shows that surplus qualifications have had a surprisingly important role in explaining the changes in wage distribution over the period 1973–2002. For example, 20 and 48 percent of the changes in the Gini coefficient of log earnings for males and females respectively can be explained by increases in the surplus qualifications factor alone (see table 3 in section 4). I also find evidence that the same factor accounts for a very significant fraction of the increase in inequality in the upper half of the distribution and has been an important determinant of why inequality did not increase as much in the lower half. This is what would be expected if over-education played the role implied by ALM and Goos and Manning.

In the next section the methodology is explained in detail. In section three I discuss the measurement of skill requirements and the processing of the Current Population Survey (CPS) earnings data. I also provide a descriptive analysis of the major trends in earnings inequality and skill mismatch. Section four presents and discusses the results. The concluding section summarizes the findings.

2 Methods

In this section I briefly explain the methods applied in the paper. First, I provide some details on the extended earnings function and the definition of the mismatch variables. I then explain how the estimated earnings function is used to decompose income inequality into factor components and the decomposition of inequality indexes based on the Shapley value.

2.1 Extended Earnings Function

Starting with Duncan and Hoffman (1981), the empirical literature on skill mismatch has been centered around the estimation of an equation of the form:

$$Y_{i,t} = X_{i,t} \cdot \gamma_t + \left[\begin{array}{ccc} Q^r_{i,t} & Q^s_{i,t} & Q^d_{i,t} \end{array} \right] \cdot \left[\begin{array}{c} \beta^r_t \\ \beta^s_t \\ \beta^d_t \end{array} \right] + \varepsilon_{i,t} \qquad (1)$$

$$E_{i,t} \equiv Q^r_{i,t} + Q^s_{i,t} - Q^d_{i,t}$$

where i and t index individuals and time respectively. $Y_{i,t}$ represents log earnings, $X_{i,t}$ is a vector of personal characteristics (including a constant and some function of age or experience), $Q_{i,t}$ represents qualifications and $\varepsilon_{i,t}$ is the error term. The vectors of parameters to be estimated are γ_t and β_t. The novelty of the approach involves splitting the education variable ($E_{i,t}$) into three parts: required (r), surplus (s), and deficit (d) qualifications.[8] The standard Mincerian approach corresponds to the particular case where $\beta^r_t = \beta^s_t = -\beta^d_t$, so

[8] All on-the-job training is assumed to be required so no decomposition applies in this case. A similar model has been estimated (Verdugo and Verdugo, 1989) that uses attained education and indicator variables for over- and under-educated workers in the right-hand-side instead of the required, deficit and surplus schooling variables. The latter model has been criticized because the returns to surplus and deficit schooling cannot be clearly identified (Cohn, 1992).

that required, surplus, and deficit qualifications all receive the same return.[9]

Given data on the individual's education attainment and the education required on the job, the surplus and deficit qualifications variables are defined as follows:

$$Q^s_{i,t} \equiv \mathbf{1}\big(E_{i,t} - Q^r_{i,t} > l\big) \cdot \big(E_{i,t} - Q^r_{i,t}\big)$$

$$Q^d_{i,t} \equiv \mathbf{1}\big(Q^r_{i,t} - E_{i,t} > l\big) \cdot \big(Q^r_{i,t} - E_{i,t}\big)$$

where $\mathbf{1}(\cdot)$ is the indicator function. $E_{i,t}$ and all qualification variables are measured in years of formal schooling. Only individuals whose education deviates at least l years from the qualifications actually required on the job are classified as mismatched. The choice of l is largely arbitrary, which means the levels of the resulting mismatch rates are relatively uninformative.[10] I set $l = 1$ but the main results are robust to different choices for this parameter within a reasonable range. Note that correctly matched individuals will have Q^r in the range $[E_{i,t} - l, E_{i,t} + l]$ and $Q^s = Q^d = 0$.

There are several extensive surveys of studies that estimate equation (1) (Green et al., 1999; Hartog, 2000; Sloane, 2003; McGuinness, 2006). As a general rule, all studies tend to confirm Sicherman's (1991) stylized facts[11] relating to the earnings of over- and under-educated workers:

[9]The other particular case of note corresponds to Thurow's (1975) job competition model, where $\beta^s_t = \beta^d_t = 0$.

[10]This is not unlike the choice of the number of weeks within which an individual must have searched for a job to be considered unemployed. There is considerable variation in the existing estimates of the incidence of skill mismatch for the U.S. and other countries. Depending on the measure utilized, the country, the period, and data source, studies have found rates of over-education ranging from 10 to 42%, with an "un-weighted" average of 23.3% in the 25 studies summarized by Groot and Maassen van den Brink (2000). Their average for under-education is 14.4%. The standard deviations are quite high: 9.9 and 8.2 percentage points respectively.

[11]Rubb (2003) provides a consistent meta-analysis of 85 estimates of the β parameters. The return to required education is 9.6% on average. Each year of

1. The earnings of over-educated workers are less than the earnings of those who have the same level of education but are in jobs where those qualifications are required (e.g. a college graduate working at a grocery store earns less on average than a college graduate who is an investment banker).

2. Over-educated workers' earnings are however generally above the earnings of workers in their same occupation or job type, who are perfectly matched qualifications-wise (i.e., the college graduate in the grocery store tends to earn more than a high-school graduate occupying a similar position).

3. The earnings of under-educated workers are more than the earnings of those with the same level of education but who are perfectly matched (e.g. a high-school graduate who becomes a manager generally earns more than the average high-school graduate).

4. The co-workers of under-educated workers who have the appropriate formal training tend to earn more than them.

A possible problem with these findings arises because of unobservable heterogeneity in individual ability. If over-educated individuals consisted of below-average ability workers then it would not be surprising to find that the returns to formal education are lower for them. While these individuals would appear to be mismatched, in reality they simply have less human capital than higher ability individuals with the same level of schooling. The converse would hold for those putatively under-educated. The hypothesis that substantial individual heterogeneity is responsible for the lower returns to surplus qualifications can be empirically tested by using panel data. Under the identifying assumption that individual ability does not vary over time, fixed effects estimates of the β parameters should effectively deal

surplus schooling yields 5.2%. Finally, deficit qualifications take away 4.8% from the required education returns.

with the problem of worker heterogeneity. Using this methodology and data from the Swedish Level of Living surveys from 1974, 1981, 1991, and 2000, Korpi and Tåhlin (2009) find that the null hypothesis of equal returns to surplus, deficit and required qualifications can be safely rejected. This study suggest that mismatch is a real issue and not just an artifice of unobservable individual heterogeneity.[12]

2.2 Decomposing Earnings Inequality

In this paper, I apply a regression-based method to decompose different inequality measures into their factor components. As a first step, I use the estimation results from equation (1) to divide log earnings into additive income components. Specifically, there are J components corresponding to the explanatory variables of the regression, one component for the constant, and one final component due to the regression residuals. Formally, let a and Z be defined as follows:

$$a_t' \equiv \left[\begin{array}{ccccc} \hat{\gamma}_t' & \hat{\beta}_t^r & \hat{\beta}_t^s & \hat{\beta}_t^d & 1 \end{array} \right]$$
$$Z_t \equiv \left[\begin{array}{ccccc} X_t & Q_t^r & Q_t^s & Q_t^d & \hat{e}_t \end{array} \right]$$

where $\hat{\gamma}$ and $\hat{\beta}$ are OLS estimates and $\hat{e}$ are residuals from equation (1). Omitting the time subscript, the income component corresponding to the j^{th} factor is $Y_j = a_j z_j$, with $Y = \sum_{j=1}^{J+2} Y_j$.

In the second step, I apply the Shapley value decomposition rule[13] to obtain each factor's contribution to earnings inequality. I study the following inequality measures: the variance, the Gini coefficient, and the 90–10, 90–50 and 50–10 percentile gaps.

[12] Other studies that estimate equation (1) using fixed effects are Bauer (2002) and Tsai (2010). Unfortunately their databases only allow voluntary transitions in and out of mismatch, which makes their results less reliable.

[13] The Shapley value decomposition is developed in Shorrocks (1999) and Sastre and Trannoy (2002). See also Israeli (2007) and Devicienti (2010).

It is easier to understand the Shapley value decomposition by first going through other simpler decomposition rules. Let $\Upsilon = \{1, 2, \ldots, J+2\}$ be the set of factor indexes. For the inequality measure $\mathrm{I}(\cdot)$ define the function:

$$\mathrm{F}\colon \{S \mid S \subseteq \Upsilon\} \to \mathbb{R} \Big/ \mathrm{F}(S) = \mathrm{I}\left(\sum_{i \in S} Y_i + \sum_{i \in \Upsilon \setminus S} \bar{Y}_i \mathbf{e}\right)$$

where $\bar{Y}_i$ is mean income from factor i and $\mathbf{e}$ is a vector of ones. The function $\mathrm{F}(S)$ gives income inequality after income from all factors not in subset S has been equalized. Clearly, $\mathrm{F}(\emptyset) = 0$ and $\mathrm{F}(\Upsilon) = \mathrm{I}(Y)$.

A desirable property of any decomposition rule is that the resulting contributions of the factors can be interpreted in an intuitively appealing way. A natural candidate is a rule that equates the contribution of each factor to its first-round marginal impact:

$$\mathrm{M}_j(\Upsilon) = \mathrm{F}(\Upsilon) - \mathrm{F}(\Upsilon \setminus \{j\}), \quad j \in \Upsilon$$

As explained in Shorrocks (1999), this decomposition rule is symmetric (or anonymous) in the sense that the contribution assigned to each factor does not depend on the way the factors are listed or labeled. However, the rule is in general not additively exact, i.e. $\sum_{j \in \Upsilon} \mathrm{M}_j(\Upsilon) \neq \mathrm{I}(Y)$.

A related decomposition rule considers the marginal impact of each factor in an elimination sequence. Let $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_{J+2})$ indicate the order in which the factors are removed and let $S(\sigma_r, \sigma) = \{\sigma_i \mid i \geq r\}$ be the set of factors that remain before factor σ_r is to be eliminated. Then the marginal impacts are given by

$$\mathrm{C}_j^{\sigma} = \mathrm{M}_j[S(j, \sigma)], \quad j \in \Upsilon$$

This decomposition rule does add up:

$$\sum_{j\in\Upsilon} \mathrm{C}_j^{\sigma} = [\mathrm{F}(\Upsilon) - \mathrm{F}(\Upsilon\backslash\{\sigma_1\})] + [\mathrm{F}(\Upsilon\backslash\{\sigma_1\}) - \mathrm{F}(\Upsilon\backslash\{\sigma_1,\sigma_2\})] + \ldots$$
$$\ldots + [\mathrm{F}(\{\sigma_{J+1},\sigma_{J+2}\}) - \mathrm{F}(\{\sigma_{J+2}\})] + [\mathrm{F}(\{\sigma_{J+2}\}) - \mathrm{F}(\emptyset)]$$
$$= \mathrm{I}(Y) - 0 = \mathrm{I}(Y)$$

There is clearly a path-dependency problem with such a rule, however. The Shapley value decomposition remedies this problem by assigning to each factor the average of its marginal impact in every possible elimination sequence. Let the set Σ contain the $(J+2)!$ possible elimination sequences. The Shapley value contribution of factor j is given by

$$\begin{aligned}\mathrm{C}_j^{Sh} =& \frac{1}{(J+2)!}\sum_{\sigma\in\Sigma}\mathrm{C}_j^{\sigma} = \frac{1}{(J+2)!}\sum_{\sigma\in\Sigma}\mathrm{M}_j[S(j,\sigma)] \qquad (2)\\ =& \sum_{s=1}^{J+2}\sum_{\substack{\{j\}\subseteq S\subseteq\Upsilon\\ |S|=s}}\frac{1}{(J+2)!}\sum_{\substack{\sigma\in\Sigma\\ S(j,\sigma)=S}}\mathrm{M}_j(S)\\ =& \sum_{s=1}^{J+2}\sum_{\substack{\{j\}\subseteq S\subseteq\Upsilon\\ |S|=s}}\frac{(J+2-s)!(s-1)!}{(J+2)!}\mathrm{M}_j(S)\end{aligned}$$

The Shapley decomposition inherits the "adding up" and "anonymity" properties from the more primitive decomposition rules it is based on. Because it considers every possible elimination sequence, it is not path-dependent. Finally and most importantly, the Shapley value has the intuitive interpretation of giving the expected marginal impact of each factor when the expectation is taken over all the possible elimination paths.

It is useful to express the Shapley decomposition in percentage terms as follows

$$\mathrm{S}_j^{Sh} \equiv \frac{\mathrm{C}_j^{Sh}}{\sum_{j\in\Upsilon}\mathrm{C}_j^{Sh}} = \frac{\mathrm{C}_j^{Sh}}{\mathrm{I}(Y)}, \quad j \in \Upsilon \tag{3}$$

Because it is an average of marginal effects over all elimination sequences, the Shapley decomposition generally depends on the level of aggregation of the factors. An important exception is the variance.[14] In this case, the marginal effect can be written:

$$\begin{aligned}\mathrm{M}_j^{\mathrm{Var}}(S) =& \mathrm{Var}\left(\sum_{i\in S} Y_i\right) - \mathrm{Var}\left(\sum_{i\in S\backslash\{j\}} Y_i\right), \quad \{j\}\subseteq S\subseteq\Upsilon \\ =& \mathrm{Var}\left(Y_j\right) + 2\mathrm{Cov}\left(Y_j, \sum_{i\in S\backslash\{j\}} Y_i\right)\end{aligned}$$

Note that the set $\Upsilon\backslash(S\backslash\{j\})$ has the same multiplier as S in the sum in equation (2). The marginal effect for this set is

$$\mathrm{M}_j^{\mathrm{Var}}\left[\Upsilon\backslash(S\backslash\{j\})\right] = \mathrm{Var}\left(Y_j\right) + 2\mathrm{Cov}\left(Y_j, \sum_{i\in\Upsilon\backslash(S\backslash\{j\})} Y_i\right)$$

It then follows that the Shapley decomposition for the variance is given by

$$\begin{aligned}\mathrm{C}_j^{Sh,\mathrm{Var}} =& \sum_{s=1}^{J+2}\sum_{\substack{\{j\}\subseteq S\subseteq\Upsilon \\ |S|=s}} \frac{(J+2-s)!(s-1)!}{(J+2)!}\frac{1}{2}\Big(\mathrm{M}_j^{\mathrm{Var}}\left[S\right] + \mathrm{M}_j^{\mathrm{Var}}\left[\Upsilon\backslash(S\backslash\{j\})\right]\Big) \\ =& \sum_{s=1}^{J+2}\sum_{\substack{\{j\}\subseteq S\subseteq\Upsilon \\ |S|=s}} \frac{(J+2-s)!(s-1)!}{(J+2)!}\left[\mathrm{Var}\left(Y_j\right) + \mathrm{Cov}\left(Y_j, \sum_{i\in\Upsilon\backslash\{j\}} Y_i\right)\right]\end{aligned}$$

[14] This applies as well to the square of the coefficient of variation.

A key point is that the expression in brackets no longer depends on S. We therefore have

$$\begin{aligned} \mathrm{C}_j^{Sh,\mathrm{Var}} =& \mathrm{Var}\big(Y_j\big) + \mathrm{Cov}\big(Y_j, Y - Y_j\big) \\ =& \mathrm{Cov}\big(Y_j, Y\big), \quad j \in \Upsilon \end{aligned} \tag{4}$$

Finally, we have

$$\mathrm{S}_j^{Sh,\mathrm{Var}} = \frac{\mathrm{Cov}\big(Y_j, Y\big)}{\mathrm{Var}(Y)}, \quad j \in \Upsilon \tag{5}$$

Equation (5) is well known in the literature on inequality decomposition. It is the decomposition into "factor inequality weights" suggested in Fields (2003). Shorrocks (1982) showed that this formula –which he calls the "natural" decomposition of the variance– is the only one satisfying a set of desirable properties for the family of continuous and symmetric inequality measures that are equal to zero if and only if income is equally distributed among all individuals. The Fields-Shorrocks decomposition rule adds-up and is also symmetric/anonymous. In addition, it is independent of the level of disaggregation. The amount of inequality accounted for by any one factor does not depend on how the other factors are grouped. Finally, it is easy to show that the J inequality weights corresponding to the regressors will add up to the R^2 of the regression.[15]

A major drawback of the Fields-Shorrocks decomposition is that it does not in general have an intuitive interpretation. A statement such as "twenty percent of earnings inequality is due to differences in education levels in the population" is generally interpreted to mean that if everyone's attained education were the same, inequality would be reduced by something close to a fifth of its original level. The decomposition into factor inequality weights only allows this kind of interpretation when inequality is measured by the variance or the

[15]See Shorrocks (1982); Fields (2003) for a complete formal statement of the properties satisfied by this decomposition rule.

square of the coefficient of variation. In contrast, the Shapley value provides an intuitively interpretable decomposition rule that varies numerically according to the inequality measure under consideration. In the case of the variance both decomposition rules coincide, and therefore the Shapley value is in this case also independent of the level of disaggregation of the factors.

One final point involves the use of Shapley values to decompose *changes* in inequality indices over time. The change in inequality index $\mathrm{I}(\cdot)$ can be written

$$\boldsymbol{\Delta}\mathrm{I}[Y(t)] = \mathrm{I}[Y(t+1)] - \mathrm{I}[Y(t)] = \sum_{j=1}^{J+2} \left[\mathrm{C}_{j,t+1}^{Sh,\mathrm{I}} - \mathrm{C}_{j,t}^{Sh,\mathrm{I}} \right]$$

The percent contribution of factor j to the change in inequality is given by:

$$\Lambda_{j,t}^{Sh,\mathrm{I}} = \frac{\mathrm{C}_{j,t+1}^{Sh,\mathrm{I}} - \mathrm{C}_{j,t}^{Sh,\mathrm{I}}}{\boldsymbol{\Delta}\mathrm{I}[Y(t)]}, \quad j \in \Upsilon \tag{6}$$

I refer to the Λ_j as differential Shapley weights. The decomposition in equation (6) has the same properties as the Shapley value decomposition for the level of inequality at a point in time.

2.3 The Yun decomposition

When inequality is measured by the variance, Yun (2006) has shown that it is possible to further decompose the changes in inequality into a price, a quantity, and a residual effect.[16]

Let the counterfactual wage distribution that would have prevailed in year $t+1$ if prices had been those of year t be defined as

$$Y_C \equiv Z_{t+1} \cdot a_t$$

[16] See also Simón (2010) for an application of this method.

Let $\mathrm{C}_{j,C}^{Sh,\mathrm{Var}}$ be the Shapley contribution to the variance of factor j under the counterfactual distribution of income. The change in the variance can be decomposed as follows

$$\begin{aligned}\boldsymbol{\Delta}\mathrm{Var}[Y(t)] =& \big(\mathrm{Var}[Y(t+1)] - \mathrm{Var}[Y_C]\big) + \big(\mathrm{Var}[Y_C] - \mathrm{Var}[Y(t)]\big) \\ =& \sum_{j=1}^{J+2}\Big[\mathrm{C}_{j,t+1}^{Sh,\mathrm{Var}} - \mathrm{C}_{j,C}^{Sh,\mathrm{Var}}\Big] + \sum_{j=1}^{J+2}\Big[\mathrm{C}_{j,C}^{Sh,\mathrm{Var}} - \mathrm{C}_{j,t}^{Sh,\mathrm{Var}}\Big]\end{aligned}$$

Noting that $\mathrm{C}_{J+2,t+1}^{Sh,\mathrm{Var}} = \mathrm{C}_{J+2,C}^{Sh,\mathrm{Var}} = \mathrm{Var}(\hat{e}_{t+1})$ we get

$$\begin{aligned}\boldsymbol{\Delta}\mathrm{Var}[Y(t)] =& \sum_{j=1}^{J+1}\Big[\mathrm{C}_{j,t+1}^{Sh,\mathrm{Var}} - \mathrm{C}_{j,C}^{Sh,\mathrm{Var}}\Big] + \sum_{j=1}^{J+1}\Big[\mathrm{C}_{j,C}^{Sh,\mathrm{Var}} - \mathrm{C}_{j,t}^{Sh,\mathrm{Var}}\Big] \\ &+ \Big[\mathrm{Var}(\hat{e}_{t+1}) - \mathrm{Var}(\hat{e}_t)\Big] \qquad (7)\end{aligned}$$

where the first, second and third terms on the right-hand-side represent, respectively, the price, characteristics and residual effects.

3 Measurement issues

In this section, I describe how the qualifications variables are constructed and briefly describe the data sources utilized. I also present a descriptive analysis of the prevalence of over- and under-education.

3.1 Skill Requirements Measure

There is consensus regarding the difficulty of measuring skill requirements. Researchers have used three main approaches, all of which have advantages and drawbacks.[17] In the present study skill requirements are measured using the job-analysis method. This measure relies on

[17] A discussion of the three methods and their comparative advantages and disadvantages can be found in Green et al. (1999) and Chevalier (2003).

systematic evaluation by professional job analysts who specify the required level of skills for the job titles in an occupational classification. In the United States this information is available in the Dictionary of Occupational Titles (DOT, U.S. Department of Labor, 1977, 1991). The DOT has clear definitions and detailed measurement instructions that all analysts are supposed to follow. Information for each of the more than 12,000 job titles is gathered through visits by Department of Labor examiners to at least two establishments in separate regions of the U.S. that employ workers in that category. They gather information on 44 different objective and subjective dimensions, including training times, required cognitive, interactive and motor skills and essential worker aptitudes, temperaments, and interests.

The most often used measures of required qualifications are called "General Educational Development" (GED).[18] On a scale of one to six, the three GED indexes measure mathematical, language and reasoning skills for each job title. Howell and Wolff (1991) analyzed the trends in the GED indexes and other DOT measures of required qualifications (1977 edition) and found that GED is highly correlated with specific vocational preparation (training time requirements), data (synthesizing, coordinating, analyzing), and three required worker aptitudes (intelligence, verbal and numerical). The GED was also correlated with a measure of interactive skills and very weakly correlated to the motor skill requirements.

Unfortunately, the DOT data collection effort is expensive, so the data is available at very low frequencies. The fourth edition of 1977 and revised fourth edition of 1991 are the last two data points.[19] Later editions of the DOT do not completely renovate the data. Rather, new editions focus on the job titles which according to the criteria of the Department of Labor experts were more likely to have undergone

[18]For example, ALM use the GED-math index as a measure of non-routine cognitive skills requirement.

[19]The Department of Labor has officially discontinued the DOT and replaced it with a new, incompatible, system called O*NET (Peterson et al., 2001, see).

significant changes.[20]

Because the DOT job title coding is not generally available in the CPS earnings files, it is necessary to aggregate the GED measure to the census 3-digit occupation level. ALM used an April 1971 CPS monthly file issued by the National Academy of Sciences in which experts assigned job title codes to each of 60,441 workers to calculate weighted sample means of the skill measures from the the DOT 1977 edition for each of the 411 occupations in the 3-digit 1970 occupational classification. Independent averages for males and females are available, so the problem generated by the heterogeneity of jobs and requirements within occupations is at least partially taken care of. To obtain averages for the 1980 classification, they applied a similar procedure to a 1980 census sample prepared for the Committee on Occupational Classification and Analysis. They also compiled averages for the 1991 Revised Fourth edition of the DOT.[21]

I merged the GED scores to the CPS data for the years 1977 and 1991 respectively. The highest of the three GED scores is the binding requirement, so I drop the other two.[22] For years other than 1977 and 1991, I let the within occupation GED scores evolve following a linear trend.[23] A final obstacle involves converting the GED score into the "years of education" unit of measurement. The GED scores are designed to be mapped into education levels. The lowest GED score corresponds to skills obtained in primary school (eg. adding and subtracting 2-digit numbers). Mid level scores require skills such as computing discounts that are normally obtained in middle and high-school. The highest GED level involves complex operations such as

[20]Spenner (1985) reviews the quality of this type of skill requirement assessment.

[21]Prof. Autor, Levy and Murnane have generously made these data publicly available.

[22]This methodological choice is unlikely to affect any results since the three scores are highly correlated (all pairwise correlations are above 0.9).

[23]Most time series variation in GED levels results from the evolution of the occupational distribution. The findings reported below are qualitatively identical if I restrict the analysis to changes between 1977 and 1991.

the analysis of dynamic systems. Using a separate dataset containing both the DOT measures and self-reported education requirements, Vaisey (2006) found that the functional form that best maps GED scores into the education requirements variable is a cubic polynomial. I follow the same approach.[24]

3.2 CPS Data

With the exception of the skill requirements measure, the data come from the NBER extracts of the CPS earnings files. The 1970 and 1980 occupational classifications necessary to merge the DOT data are available in the CPS files for the period 1973–2002. During 1973–78 earnings related questions were asked to the full CPS sample only in May. Starting in 1979, earnings questions have been asked every month to around a fourth of the sample (the outgoing rotation groups (ORG) in CPS jargon). Details on the treatment of the CPS data are discussed in the appendix. Here I only briefly discuss how the May and ORG earning supplements are processed.

As in most other studies of earnings inequality, the sample is restricted to employed wage and salary workers. Only individuals between 16 and 64 years of age with positive potential experience are kept. In trying to cope with the high non-response rates for the earnings module, starting in 1979 the BLS has allocated earnings to non-respondents by means of a hot-deck imputation method. Because earnings were not allocated to non-respondents during 1973-78, observations with imputed earnings have to be ignored to keep the series consistent over the whole period. I also drop observations for 1994 and the first eight months of 1995, a period during which allocation flags are not available.

The earnings variable we use is constructed to represent real hourly

[24]Reassuringly, the resulting average years of education required almost always coincides with the mean years actually possessed by those employed in the occupation.

earnings including overtime, tips and commissions. A known advantage of the May/ORG CPS earnings data is that it provides a point-in-time measure of earnings. Hourly earnings are weekly earnings including overtime, tips and commissions divided by usual weekly hours, except in the case when a separate (and higher) hourly rate is provided. Earnings are deflated using the CPI-U-X1 series. As in most of the literature on earnings inequality, I multiply the sampling weights by usual weekly hours so as to make the sample of hourly earnings representative of the total hours worked in the economy. I also adjust topcoded earnings, multiplying them by 1.4. After the 1994 CPS overhaul respondents with variable hours are allowed to answer that their "[weekly] hours vary". I use a method developed by Schmitt (2003) to allocate weekly hours to these workers.

The educational attainment variable is also of great importance in this study. In 1992 the education item in the CPS questionnaire was modified. Previously individuals had been asked for the highest completed grade of schooling (in years). The new item asks for the highest degree obtained. In 1998 a new battery of questions was added that permit determining the highest grade completed in most cases. I follow the imputation procedure developed by Jaeger (1997, 2003) to obtain a consistent measure of the highest grade completed over the whole period.[25]

Until 1982 the CPS used the industrial and occupational classification of the 1970 census. The 1980 census classifications are available during 1983–2002. Minor changes were introduced in the classifications in 1991, so we adjust the occupation variable in the years prior to the change to retain continuity.

[25]The exception is for individuals with at least some college in the years 1992-7. Details in the appendix.

3.3 Earnings Inequality

Wage inequality in the U.S. increased significantly during the three decades under analysis. As shown in figures 1 and 2, measures of overall inequality in log earnings like the Gini coefficient, the variance, and the 90–10 percentile gap increased substantially during the period.[26] The Gini coefficient, for example, increased from 0.144 in 1973 to 0.174 in 2002 for males, and from 0.161 to 0.178 for females.[27] This is a very significant change for earnings inequality, which usually moves slowly. The timing of the change is also interesting. Inequality remained practically constant –males– or decreased –females– during the 70s and then had an explosive period of growth during the first half of the 80s. The increase in inequality then slowed down until the early 90s. What happened to inequality in the last few years of the period depend on the measure of choice. These trends in overall earnings inequality are well documented in the literature (see for example Katz and Autor, 1999; Autor et al., 2008).

A quite different story can be told if one looks at inequality in the upper and the lower-tiers of the distribution separately. Focusing first on figure 1, after the calm 70s the 90–50 percentile gap increased sharply. Rather than slowing down and then stagnating like the Gini, however, the widening of the right half of the male wage distribution continued at the same pace into the 90s. In sharp contrast, the 50–10 gap decreased significantly after 1987. By 2002, inequality in the left half of the distribution was only slightly higher than in 1973. Indeed, the wage distribution for males was slightly left-skewed at the beginning of the period but significantly right-skewed at the end. Thus, rather than a complete stop to the trend toward increasing inequality in the 90s, there seems to have been a movement toward a

[26] Growth rates are calculated as log differences. For the percentile gaps, the growth rates correspond to the difference between the rates of growth of the corresponding percentile wages.

[27] Other commonly studied inequality indexes, like the Theil or the Atkinson index, followed a similar pattern.

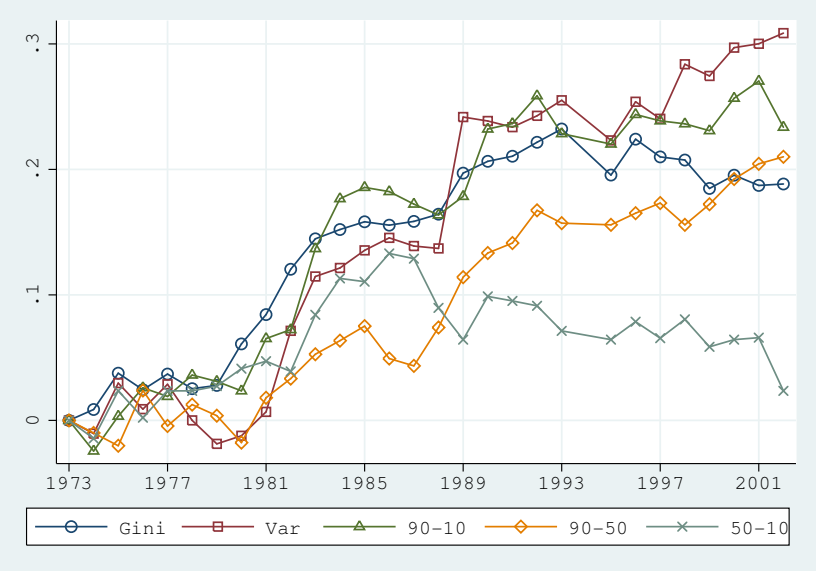

Figure 1 – The Evolution of Earnings Inequality: Males (1973=0)

polarization of earnings (Autor et al., 2008).[28] This polarizing pattern is less clear-cut for women than for men because the 50–10 gap in the female distribution did not bounce back in the 90s to the same extent.

3.4 Mismatch rates

Figure 3 shows the joint distribution of required qualifications and education at the beginning and the end of the period.[29] It is clear that workers with higher qualifications tend to be allocated to jobs with higher requirements. If workers tended to be correctly matched,

[28] Also see Lemieux (2006b), which provides evidence that over time wages have become an increasingly convex function of years of schooling.

[29] To make both years of data comparable, I use a random sub-sample of 2002 workers so that both scatter plots have roughly the same number of dots.

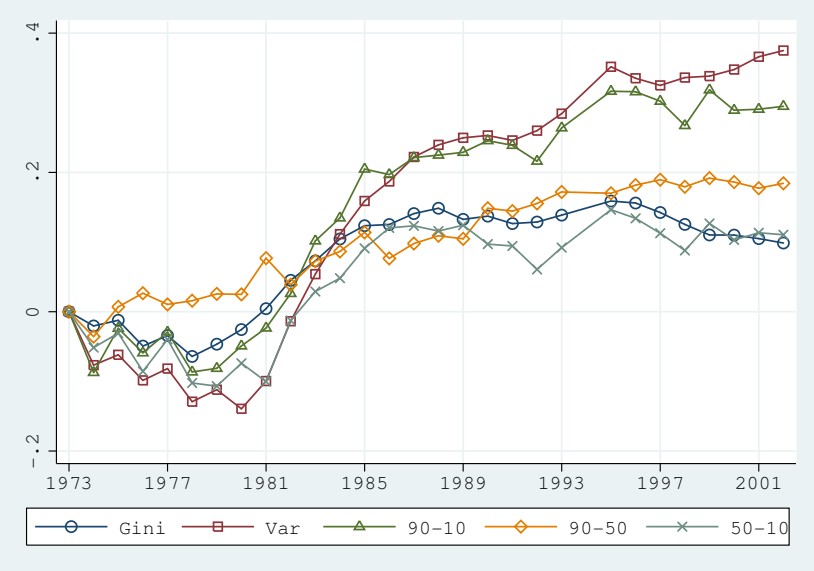

Figure 2 – The Evolution of Earnings Inequality: Females (1973=0)

the observations would be aligned along the 45 degree lines. However, the slopes from the simple OLS regressions of required qualifications on education are around 0.6.

Both for females and for males it is possible to discern two trends. First, the labor force has become more educated. Second, a much higher proportion of workers have fallen below the 45 degree line, leading to higher over-education rates. The latter point is confirmed by figure 4, which shows the evolution of mismatch rates during 1973–2002.[30] Over-education rates for males and females follow a remarkably similar path, starting in 1973 at around 15% and increasing constantly throughout the period to reach levels of around 35% of the

[30]Table 5 in the appendix contains descriptive statistics quantifying these and other trends.

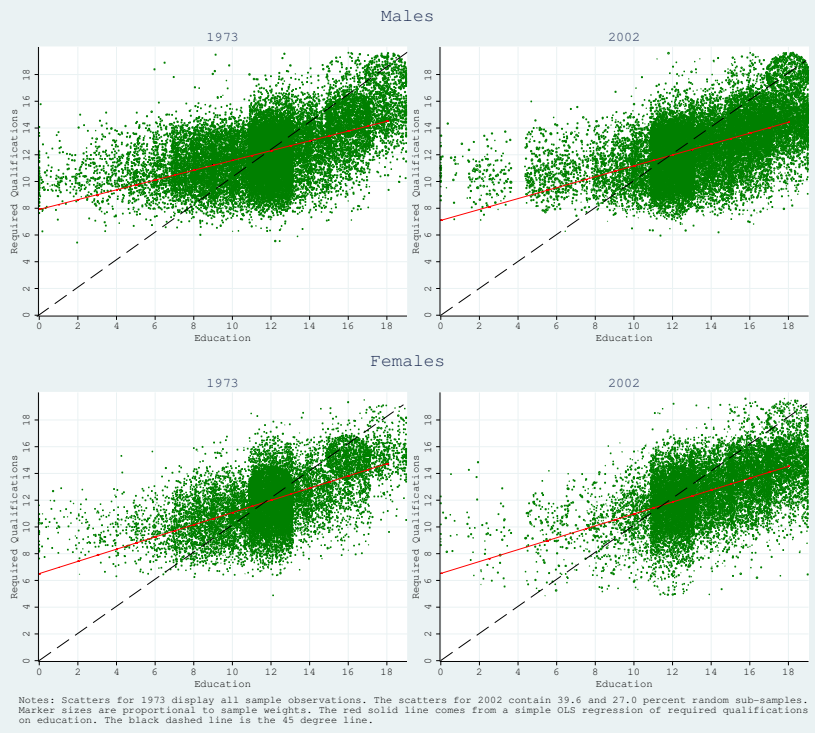

Figure 3 – Required Qualifications and Education

employed labor force. Under-education, on the contrary, follows a downward trend.

The rising over-education rate is consistent with previous analysis of the DOT and other direct measures of skill requirements (Hecker, 1992; Wolff, 2000; Handel, 2000). These studies typically show slowly rising average requirements but much faster growth in the supply of high-skill workers. However, the increasing over-education rate comes at odds with conventional thinking about recent labor market trends, specially during the 1980s. The consensus view is that relative demand for high-skill workers increased substantially during that decade, which explains why the college premium increased despite the con-

Figure 4 – Mismatch Rates 1973–2002

tinuous growth of relative supply. If demand for high-skill workers outpaced supply, how could over-education increase?

The problem with this apparent puzzle is that in the competitive model in which it is embedded skill mismatch is not possible *at all*. On one hand, in a model in which skill mismatch is a possible equilibrium outcome, obtaining a jointly increasing skill premium and over-education rate is relatively straightforward. For example, Skott (2006) obtains this result in the context of an efficiency wage model in which high-skill workers can fill both low- and high-tech jobs but low-skill workers can only be hired in low-tech positions. A negative, neutral shock to aggregate activity raises unemployment for all workers. As high- and low-skill groups compete for low-tech jobs, the relative wage in these jobs comes under pressure. As a result, there is an increase in the proportion of low-skill jobs, a rising skill-premium, and increasing over-education. The extension of this model in Slonimczyk and Skott (2010) shows that a fall in the real value of the minimum wage can produce similar results.

On the other hand, if the effect of technology on labor demand is –as argued by ALM and Goos and Manning– polarizing, then the joint occurrence of rising average returns to schooling and increasing over-

education is not puzzling. Simply put, technological change makes higher education a riskier type of investment.

4 Decomposition Results

Studies that extend the earnings function as in equation (1) typically find that the returns to required qualifications are much larger than the returns to surplus schooling and that under-educated workers are penalized for their insufficient qualifications. The differences in the returns to surplus, deficit, and required qualifications seem significant enough to motivate the suspicions that (i) skill mismatch accounts for a significant fraction of earnings inequality, and (ii) changes in mismatch rates might have contributed to the observed changes in the wage distribution. Following Fields (2003), we refer to points (i) and (ii) as the "levels" and the "differences" questions respectively.

4.1 The Levels Question

The first question can be answered by applying the *Shapley value decomposition* to different indices of earnings inequality. As explained above, a factor's Shapley value is the average marginal impact of the factor on the inequality index when all possible elimination sequences are taken into account. It is a measure of the importance of the factor in explaining earnings inequality at a point in time. The key levels question in this paper is: how important are surplus and deficit qualifications in explaining earnings inequality?

The first step in the methodology involves obtaining income components based on OLS estimation of equation (1). The results in this section are based on a specification in which the matrix of controls (X) include a full set of age dummies.[31] I have experimented with

[31]The rationale for including dummies rather than a polynomial in age is that the right functional form appears to have changed over time. A quadratic function

other reasonable specifications, with no significant change in the results.[32] For comparison purposes, I also estimate the same equations using the standard human capital specification (with actual qualifications instead of required, surplus, and deficit qualifications in the right-hand-side).

Estimation results for selected years can be found in table 6 in the appendix. The estimates are consistent with the findings of the skill mismatch literature. The returns to required qualifications are substantially higher than the returns to schooling in the standard earnings regression. Surplus qualifications yield positive but low returns, and deficit qualifications bring a penalty. The restriction $\beta^r = \beta^s = -\beta^d$ is unequivocally rejected at the 1% level of significance in all cases.[33] Both for females and males, the returns to required and surplus qualifications have increased monotonously over time, though growth was particularly strong in the 1980s.

I use these estimates to generate income components for the different regressors and the residuals.[34] In table 1, I present descriptive statistics for these components.

The sample means for the surplus and deficit qualifications income components have a straightforward interpretation. Define the premia associated with having surplus or deficit qualifications as the average difference between the log wages mismatched workers actually earn and what they would earn if they only had the qualifications that are required on their jobs (which are kept constant). Formally:

---

seems to fit well the beginning half of the series but a quartic in age seems more appropriate for later years (these changes are analyzed in detail in Lemieux, 2006b)

[32] Specifically, I experimented with: 1) a specification that allowed for non-linearities in the qualifications variables; 2) including a number of extra controls: non-white, married, industry (3 sectors), part-time, and public sector indicators, and 9 region dummies. These alternative specifications are available upon request from the author.

[33] Testing each of the two restrictions separately gave the same result.

[34] The 48 age dummies are consolidated into a single income component due to age.

Table 1 – Descriptive Statistics for Income Components (Y_j) based on the Extended Earnings Equation

A. **All Individuals**								
	Males				Females			
	1973	1983	1992	2002	1973	1983	1992	2002
$Age^{\dagger}$	0.700	0.652	0.607	0.570	0.399	0.427	0.434	0.396
	0.034	0.046	0.041	0.029	0.009	0.015	0.017	0.015
$\hat{\beta}^r Q^r$	0.896	1.064	1.406	1.545	1.187	1.332	1.644	1.674
	0.022	0.030	0.053	0.067	0.037	0.042	0.065	0.074
$\hat{\beta}^d Q^d$	-0.041	-0.026	-0.021	-0.019	-0.023	-0.012	-0.010	-0.014
	0.008	0.006	0.006	0.005	0.004	0.002	0.002	0.003
$\hat{\beta}^s Q^s$	0.014	0.022	0.044	0.059	0.020	0.032	0.055	0.077
	0.001	0.002	0.005	0.008	0.002	0.004	0.008	0.013
$\hat{e}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.171	0.175	0.184	0.202	0.142	0.141	0.160	0.177
B. **Mismatched Individuals**								
π^d	-0.190	-0.209	-0.235	-0.217	-0.162	-0.142	-0.156	-0.181
π^s	0.091	0.088	0.143	0.177	0.127	0.143	0.181	0.221

Notes: In panel A, the component's mean is in the top row and its variance in the bottom row. The constant income component is omitted. †The age income component results from the sum of the 48 age dummies. In panel B, π^s and π^d are the component's mean for overeducated and undereducated individuals respectively.

$$\pi^s \equiv \frac{1}{K} \sum_{Q^s \neq 0} \left[\left(\hat{\beta}^r Q^r + \hat{\beta}^s Q^s \right) - \hat{\beta}^r Q^r \right] = \hat{\beta}^s \cdot \overrightarrow{Q^s}$$

$$\pi^d \equiv \frac{1}{H} \sum_{Q^d \neq 0} \left[\left(\hat{\beta}^r Q^r + \hat{\beta}^d Q^d \right) - \hat{\beta}^r Q^r \right] = \hat{\beta}^d \cdot \overleftarrow{Q^d}$$

where K and H are the total counts and $\overrightarrow{x}$ and $\overleftarrow{x}$ represent the average value of x for over- and under-educated workers respectively. Note that the mismatch premia are simply the average income component for surplus and deficit qualifications for overeducated and undereducated individuals respectively.

The mismatch premia depend on the average over- and under-education depth ($\overrightarrow{Q^{s,d}}$) but *not* on over- and under-education rates. While, as already discussed, the latter changed markedly over the period, the former did not. Over/under-education depth went from 3.2/4.1 years in 1973 to 3.4/4.2 years in 2002 for males. For females, the corresponding figures were 3/3.6 years at the beginning and 3.5/3.3 years at the end of the period. As result, the mismatch premia largely followed the same path as the coefficients $\hat{\beta}^{s,d}$ in table 6. For example, during the quick rise in returns to skill in the 1980s, the over-education premium increased by 5.5 and 3.8 percentage points for males and females respectively.

The mean surplus and deficit qualifications income components can be written:

$$\hat{\beta}^s \overline{Q^s} \equiv \hat{\beta}^s \frac{1}{N} \sum_{Q^s > 0} Q_i^s = \pi^s V$$

$$\hat{\beta}^d \overline{Q^d} \equiv \hat{\beta}^d \frac{1}{N} \sum_{Q^d > 0} Q_i^d = \pi^d U$$

where N is the total number of individuals and V and U are the over- and under-education rates depicted in figure 4. The income

components for deficit and surplus qualifications are directly related to mismatch prevalence and premia. This explains why, for example, the average worker has increased average receipts of "surplus qualifications income". Both over-education rate and premia have been on the rise.

It is also possible to relate features of the distribution of the mismatch income components to the same primitive elements, but in general the expressions are not very revealing. For example, it is not hard to show that the variances can be written as:

$$\mathrm{Var}(\hat{\beta}^s Q^s) = (\hat{\beta}^s)^2 \left[V\overrightarrow{\left(Q^s\right)^2} - V^2\left(\overrightarrow{Q^s}\right)^2 \right]$$
$$\mathrm{Var}(\hat{\beta}^d Q^d) = (\hat{\beta}^d)^2 \left[U\overleftarrow{\left(Q^d\right)^2} - U^2\left(\overleftarrow{Q^d}\right)^2 \right]$$

In the empirically relevant range, the variances will be positively related to mismatch rates and depth, and to the returns to surplus and deficit qualifications. Unsurprisingly, both for females and for males, table 1 confirms a monotonic increase in the variance of the surplus qualifications income component and a fall in the variance of the penalties due to deficit qualifications.

However, it is important to emphasize that the relationship between a factor's distribution and its effect on overall income inequality is a complex one, which crucially depends on how the different components are correlated. It is not hard to imagine situations in which a factor becoming more unequally distributed leads to less overall inequality. Similarly, it is often the case that a factor contributes to inequality as measured by some indices but decreases inequality in others. The second step in the methodology addresses these issues.

In table 2, I present the Shapley value decomposition for the variance of log earnings, the Gini coefficient, and the 90–10, 90–50 and 50–10 percentile gaps. Apart from the inequality measure, the results also vary by gender, time period, and depending on whether the standard or extended earnings equation is used to decompose income. There are some striking features in the results:

Table 2 – Shapley Value Decomposition of Earnings Inequality: 1973–75, 1983–85, 1991–93, and 2000–02

	Males				Females			
	1973–75	1983–85	1991–93	2000–02	1973-75	1983–85	1991–93	2000–02
A. **VARIANCE** $\left(\mathrm{S}_j^{Sh,\mathrm{Var}}\right)$								
Mismatch Equation								
Age	15.3	19.8	16.8	12.4	6.2	9.3	8.7	7.4
Q^r	12.0	15.0	20.4	22.6	19.2	21.6	25.7	25.3
Q^d	2.5	1.8	2.2	2.3	1.6	0.7	0.9	1.4
Q^s	0.3	0.5	1.2	1.9	0.6	1.2	1.7	3.2
$\hat{e}$	69.9	62.9	59.5	60.7	72.4	67.2	63.1	62.7
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Standard Equation								
Age	16.7	21.5	18.6	13.6	6.6	10.2	9.7	8.1
E	12.0	13.1	18.2	21.7	17.8	17.8	21.5	25.4
$\hat{e}$	71.3	65.4	63.2	64.7	75.6	72.0	68.8	66.5
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
B. **GINI** $\left(\mathrm{S}_j^{Sh,\mathrm{Gini}}\right)$								
Mismatch Equation								
Age	18.4	21.9	19.0	14.6	9.3	11.8	10.9	9.8
Q^r	15.9	18.2	22.9	25.3	24.2	25.2	28.2	28.0
Q^d	4.5	2.9	2.8	2.8	2.8	1.5	1.4	1.9
Q^s	1.3	1.9	3.2	4.4	2.2	3.5	4.7	6.6
$\hat{e}$	59.8	55.1	52.1	52.8	61.5	58.1	54.8	53.7
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Standard Equation								
Age	21.0	24.7	22.0	16.9	11.1	13.9	13.2	11.7
E	17.0	17.2	21.3	25.0	22.5	22.0	25.1	28.9
$\hat{e}$	62.0	58.2	56.6	58.1	66.4	64.1	61.8	59.4
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
C. **90–10 Percentile Gap** $\left(\mathrm{S}_j^{Sh,90-10}\right)$								
Mismatch Equation								
Age	18.9	22.7	20.7	16.2	9.1	13.4	12.0	11.2
Q^r	16.6	18.7	23.7	26.3	27.4	27.3	29.9	29.4
Q^d	4.6	3.2	2.0	1.4	3.6	0.4	0.6	0.8
Q^s	1.5	1.9	3.0	4.4	2.8	3.5	3.8	6.6
$\hat{e}$	58.3	53.5	50.6	51.7	57.2	55.3	53.6	51.9
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Standard Equation								
Age	22.1	25.8	23.8	18.6	11.5	16.4	14.6	12.9
E	18.1	18.3	22.1	25.8	25.3	22.2	24.6	30.7
$\hat{e}$	59.8	55.9	54.1	55.6	63.1	61.5	60.8	56.4
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Notes: Shapley values for the constants are omitted. The income component for age is derived from 48 age dummies in the regressions.

1. The residual is the single most important explanatory factor of earnings inequality. While this fact should make us humble regarding the extent of our knowledge, it is possible to give it an optimistic interpretation. Of the innumerate factors affecting how income is distributed in a country as diverse as the United States, by simply looking at age and education we can account for a very significant fraction of the variation in earnings.

2. The residual factor is quantitatively less significant when skill mismatch is considered in the specification of the earnings equation. A simple un-weighted average of the results in the table, for example, implies that decomposing education into required, deficit and surplus qualifications reduces the importance of the residual factor by 4.3 percentage points.

3. The age factor is more important for males than for females. In contrast, the education factor is relatively more important in explaining inequality among women. When looking at mismatch factors, surplus qualifications are in almost every instance more important in explaining female wage inequality than male. In contrast, deficit qualifications tend to be more important for males.

4. Surplus and deficit qualifications jointly explain around 5% of inequality. Behind this average, however, there is substantial variation. For example, in 2000–2 surplus qualifications explained -0.7% and -1.5% of the 50–10 percentile gap for males and females respectively, meaning that this inequality index would actually increase a little if everyone were equally over-qualified. In sharp contrast, the Q^s Shapley values for the 90–50 percentile gap were 9.2% and 13.4%.

Despite the fact that only a minority of workers are mismatched and that "mismatch income" is a relatively minor income component, skill mismatch factors play a quantitatively significant role in explaining earnings inequality at any point in time. I conclude from these

Table 2 – Shapley Value Decomposition of Earnings Inequality (cont.): 1973–75, 1983–85, 1991–93, and 2000–02

	Males				Females			
	1973–75	1983–85	1991–93	2000–02	1973–75	1983–85	1991–93	2000–02
D. **90–50 Percentile Gap** $\left(\mathrm{S}_j^{Sh,90-50}\right)$								
Mismatch Equation								
Age	10.3	13.6	12.7	7.5	5.5	6.2	5.8	4.8
Q^r	20.8	21.2	24.8	26.1	32.8	27.3	28.1	26.4
Q^d	1.4	1.5	0.9	0.7	-0.2	0.7	0.4	0.3
Q^s	3.8	4.8	7.5	9.2	5.7	7.9	9.5	13.4
$\hat{e}$	63.7	59.0	54.2	56.5	56.1	58.0	56.2	55.1
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Standard Equation								
Age	10.8	13.7	11.9	7.4	5.3	6.0	4.9	5.7
E	23.0	27.2	33.4	33.9	30.0	30.3	37.5	38.3
$\hat{e}$	66.2	59.1	54.7	58.7	64.7	63.7	57.6	56.0
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
E. **50–10 Percentile Gap** $\left(\mathrm{S}_j^{Sh,50-10}\right)$								
Mismatch Equation								
Age	26.2	30.4	28.5	25.6	13.3	21.8	19.5	19.0
Q^r	13.0	16.6	22.6	26.4	21.1	27.4	32.2	33.0
Q^d	7.3	4.6	3.2	2.1	7.9	0.1	0.9	1.5
Q^s	-0.4	-0.4	-1.3	-0.7	-0.7	-1.5	-3.0	-1.5
$\hat{e}$	53.8	48.9	47.1	46.6	58.5	52.2	50.4	48.0
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Standard Equation								
Age	31.7	35.8	35.4	30.8	18.7	28.4	26.4	21.6
E	14.0	10.9	11.0	16.9	19.9	12.7	8.7	21.4
$\hat{e}$	54.3	53.2	53.5	52.3	61.4	58.8	64.9	57.0
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Notes: Shapley values for the constants are omitted. The income component for age is derived from 48 age dummies in the regressions.

results that considering what kind of jobs individuals do, and not just their qualifications, pays off in terms of explaining inequality in earnings.

4.2 The Differences Question

A noteworthy aspect of table 2 is the remarkable changes in the Shapley values that have occurred over time. The importance of the residual factor has declined, while that of required and surplus qualifications factors have generally increased. In this section I investigate whether skill mismatch factors can explain the dramatic *changes* in inequality indices over time, as shown in figures 1 and 2.

In table 3, I present a decomposition of the changes in the inequality indexes based on equation (6). In the few cases in which the change in an index was negative, a positive(negative) differential Shapley weight means that the factor exerted a force towards decreasing(increasing) inequality.[35]

One noteworthy fact is that on average the residual factor is not as important in the decomposition of inequality changes as it was found to be when analyzing the levels question. The education factor in the standard specification and the required qualifications factor in the specification with mismatch play the leading role in explaining the increase in inequality over the 30 year period. Because the importance of the residual factor is significantly smaller in the latter case, we may conclude again that the disaggregation of qualifications pays off.

[35]There are several cases in which the differential Shapley weights exceed 100 percent. These results should be interpreted to mean that the evolution of different factors exerted contradictory impulses on income inequality. To take an extreme case, between 1973 and 1983 the female 50–10 percentile gap increased 0.029 log points. Changes in the age-earnings profile and the age composition of the work force would by themselves have led to almost double such an increase. Similarly, the required qualifications factor induced an increase in inequality equal to 119 percent of the actual change. Deficit qualifications and residual income were the countervailing factors that explain why inequality did not increase more than it

Table 3 – Decomposition of Changes in Earnings Inequality

	Males				Females			
	1973 to 1983	1983 to 1992	1992 to 2002	1973 to 2002	1973 to 1983	1983 to 1992	1992 to 2002	1973 to 2002
A. **VARIANCE** $\left(\Lambda_j^{Sh,\text{Var}}\right)$								
ΔVar	0.0295	0.0372	0.0211	0.0877	0.0108	0.0470	0.0307	0.0885
Mismatch Equation								
Age	54.6	-3.5	-53.6	4.0	68.8	6.3	0.5	11.9
Q^r	34.9	64.6	54.5	52.2	49.8	47.9	26.2	40.6
Q^d	-4.9	6.0	2.5	1.5	-19.1	1.0	5.0	0.0
Q^s	1.1	7.5	12.4	6.5	7.7	5.5	12.2	8.1
$\hat{e}$	14.3	25.4	84.2	35.8	-7.2	39.2	56.0	39.4
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Standard Equation								
Age	58.1	-0.7	-59.7	4.9	78.5	8.0	0.3	13.9
E	17.4	60.4	69.0	48.0	-5.8	40.2	57.0	40.4
$\hat{e}$	24.5	40.3	90.7	47.1	27.3	51.8	42.8	45.7
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
B. **GINI** $\left(\Lambda_j^{Sh,\text{Gini}}\right)$								
ΔGini	0.0224	0.0133	-0.0059	0.0298	0.0122	0.0099	-0.0055	0.0167
Mismatch Equation								
Age	44.5	-15.5	157.4	-4.4	49.8	-4.8	33.7	22.6
Q^r	29.1	89.6	-51.1	71.8	28.4	95.1	28.0	68.2
Q^d	-8.4	0.9	3.6	-6.6	-20.3	-2.5	-13.8	-11.8
Q^s	4.5	23.4	-32.8	20.3	17.2	33.9	-47.0	48.1
$\hat{e}$	30.3	1.6	22.9	19.0	24.9	-21.6	99.2	-27.0
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Standard Equation								
Age	48.3	-8.3	180.8	-3.0	57.7	1.5	41.8	29.4
E	14.6	79.6	-86.9	63.5	-1.8	88.0	-98.5	83.2
$\hat{e}$	37.1	28.7	6.0	39.5	44.1	10.6	156.8	-12.6
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
C. **90–10 Percentile Gap** $\left(\Lambda_j^{Sh,90-10}\right)$								
Δ90–10	0.1369	0.1218	-0.0251	0.2336	0.1017	0.1144	0.0787	0.2949
Mismatch Equation								
Age	57.2	5.4	310.8	2.9	53.1	-3.2	12.4	20.4
Q^r	34.2	88.1	-119.0	78.8	16.4	64.1	20.9	36.1
Q^d	-16.5	-12.8	29.4	-19.5	-33.9	3.1	-1.7	-10.9
Q^s	5.2	19.0	-57.2	19.1	11.6	2.9	55.0	19.8
$\hat{e}$	19.9	0.4	-64.0	18.8	52.7	33.0	13.4	34.6
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Standard Equation								
Age	51.4	15.8	344.9	1.3	58.8	-13.5	13.3	18.6
E	23.4	59.9	-125.8	58.5	-21.2	60.4	119.0	47.9
$\hat{e}$	25.3	24.2	-119.1	40.2	62.4	53.1	-32.4	33.5
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Notes: The constant is omitted. The income component for age is derived from 48 age dummies in the regressions.

Table 3 – Decomposition of Changes in Earnings Inequality (cont.)

	Males				Females			
	1973 to 1983	1983 to 1992	1992 to 2002	1973 to 2002	1973 to 1983	1983 to 1992	1992 to 2002	1973 to 2002
D. **90–50 Percentile Gap** $\left(\Lambda_j^{Sh,90-50}\right)$								
Δ90–50	0.0527	0.1147	0.0427	0.2101	0.0727	0.0827	0.0289	0.1844
Mismatch Equation								
Age	49.7	12.8	-108.5	-2.6	3.6	6.2	-9.9	2.6
Q^r	13.4	48.9	55.4	41.3	-24.6	38.9	-9.6	6.2
Q^d	-9.2	-1.5	-0.2	-3.2	6.1	-4.4	-3.8	-0.2
Q^s	13.2	28.6	30.2	25.1	18.3	24.8	99.7	34.0
$\hat{e}$	32.9	11.2	123.2	39.4	96.6	34.4	23.6	57.3
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Standard Equation								
Age	45.2	3.4	-92.3	-5.6	1.1	-9.4	106.1	12.9
E	66.6	78.0	29.5	65.3	20.4	134.4	-129.7	48.0
$\hat{e}$	-11.9	18.6	162.8	40.3	78.5	-25.0	123.6	39.1
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
E. **50–10 Percentile Gap** $\left(\Lambda_j^{Sh,50-10}\right)$								
Δ50–10	0.0842	0.0071	-0.0678	0.0235	0.0290	0.0317	0.0498	0.1105
Mismatch Equation								
Age	61.9	-113.5	46.9	51.9	177.2	-27.8	25.3	49.9
Q^r	47.2	716.6	-9.2	413.3	119.4	129.9	38.6	86.0
Q^d	-21.1	-194.4	10.8	-165.5	-134.2	22.6	-0.5	-28.9
Q^s	0.1	-135.9	-2.2	-34.5	-5.2	-54.1	29.0	-3.8
$\hat{e}$	11.8	-172.9	53.8	-165.2	-57.2	29.4	7.5	-3.2
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Standard Equation								
Age	55.2	215.7	69.7	62.2	203.5	-24.4	-40.6	28.1
E	-3.7	-229.5	-28.0	-2.3	-125.6	-132.3	263.6	47.8
$\hat{e}$	48.5	113.8	58.3	40.0	22.1	256.7	-123.0	24.1
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Notes: The constant is omitted. The income component for age is derived from 48 age dummies in the regressions.

In our investigation of the levels question we found that, although significant, the role of deficit and surplus qualifications in explaining earnings inequality was modest. Also, although toward the end of the period the surplus qualifications factor appears quantitatively more important than the deficit qualifications factor, the difference is not large. The analysis of changes in inequality indexes leads to very different results. Deficit qualifications are for the most part unimportant in understanding these changes. Moreover, in most cases the distribution of this factor moved in the opposite direction, resulting in a negative –and sometimes large– differential Shapley weight for the period 1973–2002.

In sharp contrast, both for males and for females the contribution of surplus qualifications toward explaining changes in inequality is very significant. The sheer sizes of the figures for $\Lambda_{Q^s}^{Sh}$, specially for women, suggest that the over-education phenomenon is very important in understanding the changes in the wage distribution in the last three decades. For example, the decomposition exercise shows that over 20 and 48 percent of the increase in the Gini coefficient for males and females respectively can be attributed to the surplus qualifications factor.

Looking at the sub-periods, there are several interesting findings. First, note that the variance, the Gini and the 90–10 gap increased in the first two sub-periods. Required and surplus qualifications made positive contributions to these increases. The 1992–2002 sub-period is clearly different. Inequality either increased at a lower pace or decreased, mostly because of the compression of age differentials.

This same description fits well the 90–50 percentile gap but not the 50–10. The most important difference is that in the latter case surplus qualifications did not contribute to the increase in inequality of the 80s. On the contrary, according to the decomposition the increase in over-education rates and premia was one of the main fac-

---

did.

tors containing the growth of inequality in the lower half of the wage distribution.

Overall, the findings in this section suggest the surplus qualifications factor has played an important role in the changes in the wage distribution. It is one of the most important factors explaining the continuous increase in the 90–50 gap and has helped contain inequality in the 50–10 gap, the main ingredients of the polarization of income much discussed in the modern literature on earnings inequality.

4.3 The Effect of Prices and Characteristics

The differential Shapley weights provide a summary measure of the contribution of a factor toward explaining changes in inequality. In the case of the variance it is possible to further disaggregate into the price and quantity effects for each factor using equation (7).

The results for Yun's decomposition are in table 4. As expected, in every sub-period the contribution of the residual is the same as in table 3. The price effect is clearly the dominant factor explaining the rise in the variance of log earnings over the 30 year period, both for males and for females. In particular, the increase in the returns to required qualifications alone explains more than half and over a third of the change in the variance for males and females respectively.

The pattern of price effect dominating the effect of characteristics is observed in the first two subperiods. However, the results for 1992–2002 are somewhat different. For males, while the returns to required qualifications continued to increase there was a sharp contraction in age earnings differentials. At the same time, the distribution of all characteristics moved toward increasing the variance. For females, the effect of characteristics was in the same ballpark as for males, while the price effect of age and required qualifications was negligible.

The contribution of surplus qualifications to explaining changes in the variance are less impressive than for the Gini or the 90–10 percentile gap. According to the Yun decomposition, the increase

Table 4 – Yun Decomposition of Changes in the Variance of Log Earnings

	1973 to 1983			1983 to 1992			1992 to 2002			1973 to 2002		
	Price Effect	Charact Effect	Residual Effect	Price Effect	Charact Effect	Residual Effect	Price Effect	Charact Effect	Residual Effect	Price Effect	Charact Effect	Residual Effect
Males												
Age	65.7	-11.1		10.0	-13.5		-59.0	5.5		11.7	-7.7	
Q^r	36.9	-2.0		64.8	-0.2		39.8	14.7		51.7	0.5	
Q^d	1.4	-6.3		3.1	2.9		-2.7	5.2		2.5	-1.0	
Q^s	-1.4	2.5		6.0	1.5		8.7	3.7		4.7	1.8	
$\hat{e}$			14.3			25.4			84.2			35.8
Total	102.6	-17.0	14.3	83.9	-9.2	25.4	-13.3	29.1	84.2	70.7	-6.4	35.8
Females												
Age	29.7	-4.6		13.6	-7.3		-3.1	3.6		14.4	-2.5	
Q^r	25.4	-7.2		44.3	3.6		-2.4	28.6		36.0	4.6	
Q^d	-1.7	-5.3		0.5	0.5		3.9	1.1		1.3	-1.4	
Q^s	-0.4	3.2		2.9	2.7		9.9	2.3		5.3	2.8	
$\hat{e}$			-2.6			39.2			56.0			39.4
Total	53.0	-13.9	-2.6	61.3	-0.5	39.2	8.3	35.7	56.0	57.0	3.5	39.4

Notes: All entries are the percent contribution to changes in the variance over the corresponding period. Constant factor omitted.

in over-education rates was most important in the first sub-period and then decreased. In contrast, the returns to surplus qualifications started off exerting a downward pressure on the variance and ended contributing significantly to its increase.

5 Conclusions

Until a few years ago, a conventional view was that wage inequality in the U.S. has grown in time led by increases in the relative demand for high skill workers, probably due to changes in technology that favor those workers vis-a-vis the less intensively trained. The stabilization of the college premium and other inequality measures since the early 1990s —while the computer revolution is still progressing— meant that the SBTC hypothesis had to be revised. The current view is that computer technology has a complex, non-monotone relationship with skill requirements. A stylized description is that computers replace humans in routine tasks and complement humans in non-routine tasks. Depending on the skill level of the workers performing those tasks when the new technologies are introduced, they can result in demand growth that is monotone in skill (like in the 80s) or in a polarizing pattern (like in the 90s).

An important prediction of the revised SBTC perspective is that technology has a complex relationship to skill demand. While there is consensus that employment in highly desirable –"lovely" in Goos and Manning's words– jobs like managements and professional positions has increased, technological change has also resulted in fast-growing demand for service sector and other "lousy" jobs. In fact, skill requirements have not increased as fast as the early SBTC hypothesis would suggest. The DOT data presented here and in other studies shows that average skill requirements in the economy have grown very slowly during the period that elapsed between the last two editions (1977–1991). In contrast, during the period studied in the paper average levels of education attainment grew constantly and at a much

faster rate than skill requirements.

What this paper adds to the conventional view is the explicit consideration of the possibility of skill mismatch. While more educated workers tend to do relatively better in the labor market, a substantial fraction of them end up in jobs whose requirements are below their acquired levels of skill. The evidence suggests that over-qualification rates have increased substantially, while under-education has experienced a downward trend. Changes in the depth of skill mismatch, while significant, have been less impressive.

Surplus qualifications are rewarded in the marketplace to some extent. Thus, over-educated workers would be worse off had they acquired only enough education to match requirements on their jobs (assuming prices remain constant). However, they would be better off if the mismatch were eliminated through increases in requirements. The converse is true about under-educated workers. As a consequence, the contribution of the education factor toward explaining earnings inequality is more complex than what would appear at first glance.

This paper shows that skill mismatch has been a relevant cause of inequality in real earnings in the U.S. at any point in time. For example, surplus and deficit qualifications taken together account for 4.3 and 4.6 percent of the variance of log earnings, around 15 percent of the total explained variance in 2002, for males and females respectively. Moreover, the disaggregation of qualifications reduces the importance of the residual factor regardless of the inequality index under consideration.

The analysis of changes in the wage distribution shows that a substantial fraction of the increase in overall inequality during the period 1973–2002 was due to the increase in mismatch rates and mismatch premia. For example, around 20 and 48 percent of the increase in the Gini coefficients during the 30 years under analysis can be attributed to the growth in the explanatory power of surplus qualifications, again for males and females respectively.

Finally, surplus qualifications have been an important driver in the

constant growth of inequality in the upper half of the wage distribution and also important in understanding why inequality did not grow as much in the lower half, specially for males.

A Data Appendix

The main source of data are the NBER extracts of the CPS earnings files for the period 1973–2002. The CPS sample is a probability sample selected to be representative of the civilian, non-institutional population of the United States 16 years of age and older. Because of its very large size—currently about 60,000 households are interviewed each month—the CPS allows for fairly fine-grained analyses of labor market trends. An adult (the reference person) at each household is asked to report on the activities of all other persons in the household. Each household entering the CPS is administered 4 monthly interviews, then ignored for 8 months, then interviewed again for 4 more months before leaving the sample permanently. During 1973–78 earnings related questions were asked to the full CPS sample only in May. Starting in 1979, earnings questions have been asked every month to households in their fourth and last months of interviews (the outgoing rotation groups (ORG) in CPS jargon).

Sample Restrictions The study focuses on employed wage and salary workers and excludes the self-employed and those who work without pay. Only individuals between 16 and 64 years of age with positive potential experience are kept. Potential experience has the usual definition ($age - educ - 6$). A final exclusion involves individuals with allocated earnings, who could not be considered because earnings were not allocated to non-respondents during 1973–78. I also have to drop observations for 1994 and the first eight months of 1995, a period during which allocation flags are not available.

Earnings and hours The earnings variable we use is constructed to represent real hourly earnings including overtime, tips and commissions. A known advantage of the May/ORG CPS earnings data is that respondents are asked about their earnings during a reference week earlier in the month. Thus, it approximates a point-in-time measure of earnings. Our hourly earnings variable is defined as weekly earnings including overtime, tips and commissions divided by usual weekly hours, except in the case when a separate (and higher) hourly rate is provided. Topcoded earnings are multiplied by 1.4, the conventional factor adjustment to avoid bias in calculating mean earnings. Finally, due to errors at the data entry stage a small proportion of individuals have irregularly small or large weekly hours, resulting in correspondingly high or low hourly earnings. Earnings below 1 or above

100 per hour (in 1979 dollars) are therefore trimmed. We use the CPI-U-X1 series as a deflator.

The CPS has a very complex sample design, whose main purpose is to attain national and state representativeness and make sure that employment statistics are accurate. As in most of the literature on earnings inequality, I multiply the sampling weights by usual weekly hours so as to make the sample of hourly earnings representative of the total hours worked in the economy.

After the 1994 CPS overhaul respondents with variable hours are allowed to answer that their weekly "hours vary". I use a method developed by Schmitt (2003) to allocate weekly hours to these workers. Because the "hours vary" variables are not kept in the NBER files, I extract them from a set of raw CPS data files and merge them with the NBER dataset. Individuals answering their hours vary do indicate whether they work full or part-time. We use regression predicted values to impute usual hours for these individuals. Four separate regressions are used according to gender and full-time status. The predictors are a quadratic function of age, a set of race and education dummies, marital status, indicators for foreign born and US citizens, and dummies for union, public sector, manufacturing, and services. A small number of individuals who answer their hours vary provide hours worked at the reported hourly wage (typically these workers work different jobs at different rates and the interviewer records the hourly rate at the job with the largest number of hours). In this case we give priority to the latter amount—a true response—over the regression imputation.

Education The educational attainment variable is also of great importance in this study. In 1992 the education item in the CPS questionnaire was modified. Previously individuals had been asked for the highest completed grade of schooling (in years). The new item asks for the highest degree obtained. In 1998 a new battery of questions was added that permit determining the highest grade completed in most cases. I follow the imputation procedure developed by Jaeger (1997, 2003) to obtain a consistent measure of the highest grade completed over the whole period.

Unfortunately the scarcity of information during the period 1992–97 results in no individuals being imputed 15 or 17 years of education. This feature of the data leads to improbable jumps in the mismatch prevalence series in 1992 and 1997. To address this issue I first linked the 1997 and 1998 files. The extra information available in 1998 could then be used to improve the imputation method for those individuals present in both datasets and whose answers to the completed degree questions where the same in both years. This adjustment is enough to almost eliminate the jump in the series from 1997 to 1998. For the individuals in the 1997 sample whose education could not be determined in this way, and for the

respondents during the 1992–96 period I used a refinement of the Jaeger method to impute completed years of schooling to those whose answer to the completed degree question was problematic (i.e. the "some college," "college," and "advanced degree" categories). First, for each of the problematic categories I estimated separate ordered probits of completed years of schooling using the more complete 1998 survey. The predictors were sex, age, non-white, marital status, public sector, manufacturing, services and a set of dummies for the 2-digit occupation. These estimates were used to predict the probability of belonging to each of the completed years of schooling categories for the individuals in the imputation sample. The imputation decision was done according to random assignment to each of the completed years of schooling categories conditional on the predicted probabilities.

Table 5 – Descriptive Statistics

	Males				Females			
	1973	1983	1992	2002	1973	1983	1992	2002
$\ln W$	1.91	1.79	1.77	1.88	1.51	1.47	1.55	1.69
	0.49	*0.52*	*0.56*	*0.57*	*0.44*	*0.45*	*0.50*	*0.53*
E	11.90	12.85	13.23	13.50	12.01	12.88	13.42	13.79
	3.14	*2.94*	*2.94*	*3.02*	*2.65*	*2.50*	*2.59*	*2.73*
Overeduc (V)	15.6%	25.3%	31.1%	33.5%	15.5%	22.4%	30.5%	34.8%
Undereduc (U)	21.8%	12.2%	8.8%	8.9%	14.3%	8.6%	6.7%	7.7%
Q^r	12.27	12.44	12.45	12.60	11.99	12.32	12.49	12.66
	2.04	*2.03*	*2.04*	*2.11*	*1.95*	*1.89*	*1.93*	*2.06*
Q^s	0.49	0.82	1.03	1.14	0.46	0.73	1.03	1.20
	1.21	*1.51*	*1.66*	*1.73*	*1.13*	*1.44*	*1.66*	*1.78*
Q^d	0.89	0.49	0.37	0.37	0.51	0.29	0.23	0.26
	1.91	*1.48*	*1.37*	*1.34*	*1.39*	*1.05*	*0.99*	*1.00*
Age	36.93	36.17	36.77	38.51	36.73	35.64	36.96	38.98
	12.73	*12.02*	*11.14*	*11.47*	*13.32*	*12.11*	*11.19*	*11.71*
Married	79.7%	70.5%	64.6%	63.0%	65.2%	59.3%	56.2%	54.6%
Non-white	14.6%	17.4%	22.6%	28.2%	17.0%	19.2%	22.6%	27.9%
Part-time	1.6%	3.5%	3.3%	2.3%	2.7%	5.8%	4.4%	2.7%
Public Sector	15.6%	16.0%	15.0%	13.4%	22.2%	20.5%	20.5%	20.2%
Manufacturing	34.5%	29.6%	26.1%	20.9%	23.4%	18.4%	15.0%	10.6%
Services	62.8%	67.5%	71.3%	76.7%	75.9%	80.9%	84.2%	88.6%
Sample Size	23,425	76,770	72,213	59,765	16,154	68,009	69,552	59,724

Notes: Standard deviations for continuous variables are in italics under the sample means.

Table 6 – OLS Estimation Results: 1973, 1983, 1992, and 2002

	Males				Females			
	1973	1983	1992	2002	1973	1983	1992	2002
Mismatch Equation								
Q^r	0.073***	0.086***	0.113***	0.123***	0.099***	0.108***	0.132***	0.132***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
Q^d	-0.046***	-0.052***	-0.056***	-0.052***	-0.045***	-0.042***	-0.045***	-0.054***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.002)	(0.002)
Q^s	0.029***	0.027***	0.043***	0.052***	0.042***	0.044***	0.054***	0.064***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.003)	(0.001)	(0.001)	(0.001)
R^2	0.297	0.358	0.404	0.389	0.271	0.313	0.368	0.375
Var($\hat{e}$)	0.171	0.175	0.184	0.203	0.142	0.141	0.160	0.177
Obs	23,423	76,770	72,213	59,765	16,150	68,009	69,552	59,724
Standard Equation								
E	0.054***	0.059***	0.075***	0.082***	0.071***	0.073***	0.087***	0.094***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
R^2	0.283	0.335	0.366	0.349	0.242	0.267	0.307	0.336
Var($\hat{e}$)	0.174	0.181	0.196	0.215	0.148	0.151	0.175	0.188
Obs	23,425	76,770	72,213	59,765	16,154	68,009	69,552	59,724

Estimation results for 48 age dummies and the constant are omitted. Standard errors in parentheses.
***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

References

Daron Acemoglu. Technical Change, Inequality, and the Labor Market. *Journal of Economic Literature*, 40(1):7–72, 2002.

Daron Acemoglu and David Autor. Skills, Tasks and Technologies: Implications for Employment and Earnings. In Orley Ashenfelter and David E. Card, editors, *Handbook of Labor Economics*, volume 4B, chapter 12, pages 1043–1171. Elsevier Science B.V., Amsterdam, 2011.

James Albrecht and Susan Vroman. A Matching Model with Endogenous Skill Requirements. *International Economic Review*, 43(1): 283–305, 2002.

David Autor, Lawrence Katz, and Melissa Kearney. Trends in U.S. Wage Inequality: Revising the Revisionists. *The Review of Economics and Statistics*, 90(2):300–323, 2008.

David H. Autor, Frank Levy, and Richard J. Murnane. The Skill Content of Recent Technological Change: An Empirical Exploration. *The Quarterly Journal of Economics*, 118:1279–334, 2003.

Thomas K. Bauer. Educational Mismatch And Wages: A Panel Analysis. *Economics of Education Review*, 21(3):221–229, 2002.

Francine D. Blau and Lawrence M. Kahn. The U.S. Gender Pay Gap in the 1990s: Slowing Convergence. *Industrial and Labor Relations Review*, 60(1):45–66, 2006.

Lex Borghans and Andries de Grip, editors. *The Overeducated Worker? The Economics of Skill Utilization.* Edward Elgar, Northampton, MA, 2000.

John Bound and George Johnson. Changes in the Structure of Wages in the 1980's: An Evaluation of Alternative Explanations. *The American Economic Review*, 82(3):371–392, 1992.

S. Bowles, H. Gintis, and M. Osborne. The Determinants of Earnings: A Behavioral Approach. *Journal of Economic Literature*, 39(4): 1137–1176, 2001.

Felix Büchel, Andries de Grip, and Antje Mertens, editors. *Overeducation in Europe*. Edwad Elgar, Northampton, MA, 2003.

David Card and John DiNardo. Skill-Biased Technological Change and Rising Wage Inequality: Some Problems and Puzzles. *Journal of Labor Economics*, 20(4):733–783, 2002.

Arnaud Chevalier. Measuring Over-education. *Economica*, 70:509–531, 2003.

Elchanan Cohn. The Impact of Surplus Schooling on Earnings: Comment. *The Journal of Human Resources*, 27(4):679–682, 1992.

F. Devicienti. Shapley-value Decompositions of Changes in Wage Distributions: a Note. *Journal of Economic Inequality*, 8(1):35–45, 2010.

John DiNardo, Nicole Fortin, and Thomas Lemieux. Labor Market Institutions and the Distribution of Wages, 1973–1992: A Semiparametric Approach. *Econometrica*, 64(5):1001–44, 1995.

Greg J. Duncan and Saul D. Hoffman. The Incidence And Wage Effects Of Overeducation. *Economics of Education Review*, 1(1): 75–86, 0 1981.

Gary Fields. Accounting For Income Inequality and its Change: A New Method, with Application to the Distribution of Earnings in the United States. *Research in Labor Economics*, 22:1–38, 2003.

Richard B. Freeman. *The Over-educated American*. Academic Press, New York, 1976.

Maarten Goos and Alan Manning. Lousy and Lovely Jobs: The Rising Polarization of Work in Britain. *The Review of Economics and Statistics*, 89(1):118–133, Jan 2007.

Francis Green, Steven McIntosh, and Anna Vignoles. 'Overeducation' and Skills – Clarifying the Concepts. *CEP Discussion Paper*, 435, 1999.

Wim Groot and Henriëtte Maassen van den Brink. Overeducation In The Labor Market: A Meta-Analysis. *Economics of Education Review*, 19(2):149–158, 2000.

Michael J. Handel. Trends in Direct Measures of Job Skill Requirements. *Levy Institute Working Paper*, (301), 2000.

Joop Hartog. Over-Education And Earnings: Where Are We, Where Should We Go? *Economics of Education Review*, 19(2):131–147, 2000.

Daniel E. Hecker. Reconciling Conflicting Data on Jobs for College Graduates. *Monthly Labor Review*, pages 3–12, 1992.

David R. Howell and Edward N. Wolff. Trends in the Growth and Distribution of Skills in the U.S. Workplace, 1960-1985. *Industrial and Labor Relations Review*, 44(3):486–502, 1991.

O. Israeli. A Shapley-based Decomposition of the R-Square of a Linear Regression. *Journal of Economic Inequality*, 5(2):199–212, 2007.

David A. Jaeger. Reconciling the Old and New Census Bureau Education Questions: Recommendations for Researchers. *Journal of Business & Economic Statistics*, 15(4):300–9, 1997.

David A. Jaeger. Estimating the Returns to Education using the newest Current Population Survey Education Questions. *Economic Letters*, 78:385–394, 2003.

Chinhui Juhn, Kevin M. Murphy, and Brooks Pierce. Wage Inequality and the Rise in Returns to Skill. *The Journal of Political Economy*, 101(3):410–442, 1993.

Lawrence Katz and David Autor. Changes in the Wage Structure and Earnings Inequality. In Orley Ashenfelter and David Card, editors, *Handbook of Labor Economics*, volume 3A, pages 1463–1555. Elsevier Science B.V., Amsterdam, 1999.

Lawrence Katz and Kevin Murphy. Changes in Relative Wages, 1963–87: Supply and Demand Factors. *The Quarterly Journal of Economics*, (107):35–78, 1992.

T. Korpi and M. Tåhlin. Educational Mismatch, Wages, and Wage Growth: Overeducation in Sweden, 1974–2000. *Labour Economics*, 16(2):183–193, 2009.

David Lee. Wage Inequality in the US During the 1980s: Rising Dispersion or Falling Minimum Wage? *Quarterly Journal of Economics*, (114):941–1024, 1999.

Thomas Lemieux. Increasing Residual Wage Inequality: Composition Effects, Noisy Data, or Rising Demand for Skill? *The American Economic Review*, 96(3):461–498, 2006a.

Thomas Lemieux. The Mincer Equation Thirty Years after Schooling, Experience, and Earnings. In S. Grossbard-Shechtman, editor, *Jacob Mincer, A Pioneer of Modern Labor Economics*, chapter 11. Springer Verlag, 2006b.

Frank Levy and Richard Murnane. U.S. Earning Levels and Earnings Inequality: A Review of Recent Trends and Proposed Explanations. *Journal of Economic Literature*, 30(3):1333–1381, Sep 1992.

Séamus McGuinness. Overeducation in the Labour Market. *Journal of Economic Surveys*, 20(3):387–418, 2006.

Kevin Murphy and Finis Welch. The Structure of Wages. *The Quarterly Journal of Economics*, 107(1):285–326, 1992.

N. Peterson, M. Mumford, W. Borman, R. Jeanneret, E. Fleishman, K. Levin, M. Campion, M. Mayfield, F. Morgeson, K Pearlman,

M. Gowing, A. Lancaster, M. Silver, and D. Dye. Understanding Work Using The Occupational Information Network (O*Net): Implications For Practice And Research. *Personnel Psychology*, 54(2): 451–492, 2001.

S. Rubb. Overeducation In The Labor Market: A Comment And Re-Analysis Of A Meta-Analysis. *Economics of Education Review*, 22 (6):621–629, 2003.

M. Sastre and A. Trannoy. Shapley Inequality Decomposition by Factor Components: Some Methodological Issues. *Journal of Economics*, 9:51–89, 2002.

Michael Sattinger. Assignment Models of the Distribution of Earnings. *Journal of Economic Literature*, 31(2):831–880, 1993.

John Schmitt. Creating a Consistent Hourly Wage Series from the Current Population Survey's Outgoing Rotation Group, 1979–2002. *CEPR Working Paper*, 2003.

A. F. Shorrocks. Inequality Decomposition by Factor Components. *Econometrica*, 50(1):193–211, 1982.

A.F. Shorrocks. Decomposition Procedures for Distributional Analysis: a Unified Framework Based on the Shapley Value. *Mimeo University of Essex*, 1999.

Nachum Sicherman. "Overeducation" in the Labor Market. *Journal of Labor Economics*, 9(2):101–122, 1991.

H. Simón. International Differences in Wage Inequality: a New Glance with European Matched Employer–Employee Data. *British Journal of Industrial Relations*, 48(2):310–346, 2010.

Peter Skott. Wage Inequality and Overeducation in a Model with Efficiency Wages. *Canadian Journal of Economics*, 39(1):94–123, 2006.

Peter J. Sloane. Much Ado About Nothing? What Does the Overeducation Literature Really Tell Us? In Büchel et al. (2003), chapter 2, pages 11–48.

Fabian Slonimczyk and Peter Skott. Employment and Distribution Effects of the Minimum Wage. *University of Massachusetts Working Paper*, (3), 2010.

Kenneth I. Spenner. The Upgrading and Downgrading of Occupations: Issues, Evidence, and Implications for Education. *Review of Educational Research*, 55(2):125–154, 1985.

Lester Thurow. *Generating Inequality*. Basic Books, New York, 1975.

Yuping Tsai. Returns to Overeducation: A Longitudinal Analysis in the U.S. Labor Market. *Economics of Education Review*, 29(4): 606–617, 2010.

U.S. Department of Labor. *Dictionary of Occupational Titles*. Government Printing Office, 4th edition, 1977.

U.S. Department of Labor. *Dictionary of Occupational Titles*. Government Printing Office, revised 4th edition, 1991.

Stephen Vaisey. Education and Its Discontents: Overqualification in America, 1972–2002. *Social Forces*, 85(2):835–864, 2006.

Richard R. Verdugo and Naomi Turner Verdugo. The Impact of Surplus Schooling on Earnings: Some Additional Findings. *The Journal of Human Resources*, 24(4):629–643, 1989.

Edward Wolff. Technology and Demand for Skills. In Borghans and de Grip (2000), chapter 2, pages 27–56.

M.S. Yun. Earnings Inequality in USA, 1969–99: Comparing Inequality Using Earnings Equations. *Review of Income and Wealth*, 52 (1):127–144, 2006.